MONTE CARLO METHODS FOR POSTERIOR DISTRIBUTIONS ASSOCIATED WITH MULTIVARIATE STUDENT'S t DATA

By

DOBRIN MARCHEV

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2004

Copyright 2004

by

Dobrin Marchev

To the memory of my grandfather, Dimiter

ACKNOWLEDGMENTS

I would like to thank Jim Hobert for his support throughout the last 3 years. His guidance, endless questions and rigorous criticism brought the necessary quality to my work. Without his insights, this dissertation would have never been the same.

I would also like to thank Brett Presnell, Jim Booth, Alex Trindade, Murali Rao and Mike Daniels for serving on my committee and spending their valuable time with my dissertation. To Brett Presnell, I am particularly thankful for stimulating discussions throughout my studies at the University of Florida. I learned a lot about probability, mathematics, and LaTex from him.

Above all, I thank Pavlina for her love and for sharing with me all the moments of happiness and grief.

Finally, I thank my parents, Angel and Liliana, who supported me and pointed me in the right direction for the first 25 years of my life.

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

ABSTRACT ix

CHAPTER

1 INTRODUCTION 1

1.1 From the Normal to the Student's t Model 1
1.2 Outline 4
1.3 How to Sample from an Intractable Density 4

2 MONTE CARLO METHODS FOR THE STUDENT'S t MODEL 7

2.1 The Model 7
2.1.1 Definition 7
2.1.2 Propriety of the Posterior 8
2.2 Making Exact Draws from the Target Posterior 13
2.3 Simulation Examples 24

3 DATA-AUGMENTATION ALGORITHMS 26

3.1 Introduction 26
3.2 Standard Data Augmentation 27
3.3 Meng and van Dyk's Extensions 30
3.3.1 Conditional Augmentation 31
3.3.2 Marginal Augmentation 32
3.3.3 Marginal Augmentation with Improper Prior 34
3.4 Connection with Group Theory 39
3.5 Understanding the Artwork of van Dyk and Meng 47
3.6 Other Problems with van Dyk and Meng's Lemma 1 51

4 GEOMETRIC ERGODICITY OF $\Phi^{\star}_{\mathrm{MA}}$ FOR THE MULTIVARIATE STUDENT'S t PROBLEM 55

4.1 Introduction 55
4.2 Markov Chain Background 56

4.3 Multivariate Student's t Problem 59
4.4 Numerical Example 66
4.5 Concluding Remarks 70

REFERENCES 72

BIOGRAPHICAL SKETCH 76

LIST OF TABLES

Table page

2–1 Acceptance rates of the rejection sampler when $d = 2$ 25

2–2 Acceptance rates of the rejection sampler when $d = 3$ 25

LIST OF FIGURES

Figure page

4–1 Empirical comparison of Markov chains and the target 69

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

MONTE CARLO METHODS FOR POSTERIOR DISTRIBUTIONS
ASSOCIATED WITH
MULTIVARIATE STUDENT'S t DATA

By

Dobrin Marchev

August 2004

Chair: James P. Hobert
Major Department: Statistics

Let f denote the posterior density that results when a random sample of size n from a d-dimensional location-scale Student's t distribution (with ν degrees of freedom) is combined with the standard non-informative prior. We consider several Monte Carlo methods for sampling from the intractable density f, including rejection samplers and Gibbs samplers. Special attention is paid to the Markov chain Monte Carlo (MCMC) algorithm developed by van Dyk and Meng who provided considerable empirical evidence suggesting that their algorithm converges to stationarity much faster than the standard data-augmentation Gibbs sampler. In addition to its practical importance, this algorithm is interesting from a theoretical standpoint because it is based on a Markov chain that is *not positive recurrent.* We formally analyze the relevant marginal chain underlying van Dyk and Meng's algorithm. In particular, we establish drift and minorization conditions showing that, for many (d, ν, n) triples, the marginal chain is geometrically ergodic. This is the first general, rigorous analysis of an MCMC algorithm based on a non-positive recurrent Markov chain. Moreover, our results are important from a practical

standpoint since geometric ergodicity guarantees the existence of central limit theorems that can be used to calculate Monte Carlo standard errors; and because the drift and minorization conditions themselves allow for the calculation of exact upper bounds on the total variation distance to stationarity. Our results are illustrated using several numerical examples.

CHAPTER 1
INTRODUCTION

1.1 From the Normal to the Student's t Model

Scientists have built all kinds of models to explain the surrounding world. Statisticians, in particular, use stochastic models in an attempt to describe and summarize data; and to predict future observations. The most widely used statistical models are based on the assumption that the data are normally distributed. For example, a very simple model might assume that the sample data, $y_1, \ldots, y_n$, are realizations of independent and identically distributed (iid) normal random variables. That is,

$$y_i \sim \mathrm{N}(\mu, \sigma),\ i = 1, \ldots, n, \tag{1.1}$$

where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+ := (0, \infty)$ are the unknown parameters denoting mean and variance, respectively. To accommodate more complex sampling designs, the model in Equation (1.1) can be generalized in many ways. For example, it can be extended to multivariate scenarios allowing y_is to be vectors (of possibly different dimensions), or it can be assumed that the mean μ is dependent on other variables (regression analysis). One reason these models have been studied so extensively is that they can be analyzed relatively easily.

Unfortunately, there are situations where the normal distribution is inappropriate because of its light tails. A possible solution to this problem is to change the model in a different direction. Instead of the normal distribution, one can use a distribution with heavier tails. A natural candidate is the Student's t distribution.

Therefore, can replace the assumption in Equation (1.1) with

$$y_i \sim t(\nu, \mu, \sigma),\ i = 1, \ldots, n, \tag{1.2}$$

where $t(\nu, \mu, \sigma)$ is the location-scale Student's t distribution (Section 2.1.1) and ν is the degrees of freedom (assumed to be known). This idea is not new. Lange, Little and Taylor (1989) extensively listed articles and books using different t models, claiming that the first use dates back at least to Jeffreys (1939), who fit it to a series of astronomical data. Of course, the assumption in Equation (1.2) can also be generalized in many ways. Lange et al. (1989) described results from applying different generalizations of (1.2) to all sorts of real data sets, ranging from the stack-loss data set of Brownlee (1965) to the traffic-accident data of Draper and Smith (1981, p. 191). Another modification of the model in Equation (1.2) is given by Zellner (1976) who considers $\boldsymbol{y} := (y_1, \ldots, y_n)^\top$ to be multivariate t distributed with identity covariance matrix, instead of assuming that the individual t_is are independently t distributed. Notice that, contrary to the normal case, these are not equivalent models, as it is suggested by Kelejian and Prucha (1985).

Heavier-tailed distributions are especially needed when modeling the high volatility inherent in financial data sets (see; e.g., Mandelbroit, 1963). In particular, the *market model* is a multivariate linear regression extension of (1.2) that relates the returns on assets ($\boldsymbol{y}$) to an overall market return (embedded in μ). Substantial empirical evidence shows that the market model error distribution should be heavier-tailed than the multivariate normal (see; e.g., Affleck-Grave and McDonald, 1989). Praetz (1972) considers a similar model that describes the change in the log of several share-price index series with a t distribution, and compares the fit (by χ^2-test) to three other models, concluding that *only* the t model is acceptable, even at 0.01 level of significance. Many studies concerned with economical or financial applications, like Fernández and Steel (2000), use Bayesian t models, changing the

model of Equation (1.2) to

$$\begin{aligned} y_i|\mu,\sigma &\sim t(\nu,\mu,\sigma),\ i=1,\ldots,n; \\ \mu,\sigma &\sim \pi(\mu,\sigma), \end{aligned} \tag{1.3}$$

where $\pi(\mu,\sigma)$ is a (possibly improper) prior. The multivariate version of (1.3) is the focus of our attention. We assume that given $(\boldsymbol{\mu},\boldsymbol{\Sigma})$[1], the vectors $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ are conditionally independent with

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{\mu},\boldsymbol{\Sigma} &\overset{\text{iid}}{\sim} t_d(\nu,\boldsymbol{\mu},\boldsymbol{\Sigma}),\ i=1,\ldots,n; \\ \boldsymbol{\mu},\boldsymbol{\Sigma} &\sim \pi(\boldsymbol{\mu},\boldsymbol{\Sigma}), \end{aligned} \tag{1.4}$$

where $t_d(\nu,\boldsymbol{\mu},\boldsymbol{\Sigma})$ represents the d-dimensional Student's t distribution (whose density is defined in Section 4.3) and $\pi(\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{d+1}{2}}$ is the standard non-informative prior for a multivariate location-scale problem.

Even though the Student's t models are more realistic, there is a practical limitation to their use: making inferences about μ and σ is not as straightforward as in the normal case. For example, the maximum likelihood estimators of μ and σ in Equation (1.2) are not available in closed form. Similarly, there is no known exact method to sample directly from the posterior distribution resulting from Equations (1.3) and (1.4). A possible solution to these problems is to apply the EM algorithm to (1.2) or a Gibbs sampler to (1.4). We concentrate on the Bayesian paradigm and give some answers to the question of how to sample from the posterior distribution associated with Equation (1.3).

Geweke (1993) was among the first to study a Bayesian t model with independent errors, but a mistake in the main result invalidates his proof of the propriety

[1] Throughout the dissertation, we use bold for variables that are vectors or matrices

of the posterior. Geweke (1993) was also among the first to suggest a Gibbs sampler to get draws from the posterior and to actually apply his results to a real data set involving U.S. macroeconomic time series, studied by Nelson and Plosser (1982). van Dyk and Meng (2001) and Meng and van Dyk (1999) show how to improve on the standard Gibbs method for sampling from the t posterior, but they do not deal with the problem of the propriety of the posterior.

1.2 Outline

The purpose of this dissertation is to present and compare different Monte Carlo methods for sampling from the posterior resulting from Equation (1.4). The organization is as follows.

In Chapter 2 we define the most general version of the Student's t model that will be considered herein, and we prove a result that guarantees the propriety of the posterior. We then show how to use classical Monte Carlo methods in order to sample from the posterior, and illustrate with some examples.

Chapter 3 deals with MCMC methods of getting draws from the target posterior associated with our model. We first introduce the standard data-augmentation algorithm, and then we attempt to explain the new MCMC methods developed by Meng and van Dyk (1999), van Dyk and Meng (2001), and Liu and Wu (1999). In our view, these papers did not provide a rigorous foundation for these new methods. In this chapter we attempt to lay such a foundation.

Finally, in Chapter 4, we state and prove our results about the geometric ergodicity of van Dyk and Meng's algorithm.

Before moving to Chapter 2, we conclude this chapter with a brief description of various methods for obtaining draws from an intractable density.

1.3 How to Sample from an Intractable Density

Suppose we are given a density, $f(\theta)$, where $\theta \in \Theta$. We may be interested in computing a particular numerical characteristic of f, like its mean or standard

deviation; but sometimes, the complexity of the expressions may not allow us to do the computations directly. In such cases, there are several methods for estimating or approximating the properties of interest of f. These include asymptotic approximations, numerical integration, and sampling. Sampling methods (also called Monte Carlo methods) are ways of generating samples from a density that equals f or that approximates f. Using such samples, we can estimate the characteristics of the target density, f. For example, if we denote the sample by $\theta_0, \theta_1, \ldots$, then there are strong laws of large numbers that guarantee (under some regularity conditions) that

$$\frac{1}{n}\sum_{i=0}^{n-1} g(\theta_i) \rightarrow \int_{\Theta} g(\theta) f(\theta) d\theta \text{ a.s.},$$

as $n \rightarrow \infty$ for any integrable function $g : \Theta \rightarrow \mathbb{R}$.

There are many different sampling methods. In some cases, it is possible to sample directly from f, and thus obtain an independent and identically distributed (iid) sample. When direct sampling is infeasible (especially when we don't know the normalizing constant of f), a possible way to get iid draws from the target density f is the *rejection sampler*. Here is a brief description of this method. Suppose we know how to sample from the density $h(\theta), \theta \in \Theta$ (often called a *candidate*) and we also know that $\frac{f(\theta)}{h(\theta)} \leq M < \infty \; \forall \; \theta \in \Theta$. Then the basic algorithm of the rejection sampler consists of performing the following steps:

1. Generate θ from h
2. Generate u from $\mathrm{U}(0, 1)$
3. If $u\, M\, h(\theta) \leq f(\theta)$, then return θ, or else go to 1.

More details about this algorithm, along with some examples and a proof that it returns iid samples from f, can be found in Ripley (1987).

In many problems it is impractical or even impossible to get an iid sample, and then Markov chain Monte Carlo (MCMC) is often used. Generally speaking, when an MCMC algorithm is applied, dependent samples are generated using a Markov

chain with stationary density f. The most commonly used MCMC algorithms are the Metropolis-Hastings algorithm and the Gibbs sampler, which itself can be viewed as a special case of the Metropolis algorithm. Tierney (1994) gives a good overview of MCMC methods for exploring intractable densities.

In this dissertation we consider the special case when f is the posterior density that results when data from a multivariate location-scale Student's t distribution are combined with the standard non-informative prior. van Dyk and Meng (2001) developed an efficient MCMC algorithm for sampling from f, and provided empirical evidence suggesting that their algorithm converges to stationarity much faster than the standard Gibbs sampler. We formally analyze the Markov chain underlying Meng and van Dyk's algorithm. Specifically, we derive drift and minorization conditions, which can be used to get rigorous bounds on the total variation distance to stationarity, and to do regenerative simulation.

CHAPTER 2
MONTE CARLO METHODS FOR THE STUDENT'S t MODEL

2.1 The Model

In many statistical models, the observations are assumed to be normally distributed. For example, inference procedures for the classical linear regression and ANOVA models are based on the normality assumption. In many applications, however, there is a need for heavier-tailed distributions. In particular, observations on daily returns of stocks (and financial data in general) prove to be incompatible with the assumption of normality. The Student's t model is a relatively tractable heavy-tailed alternative.

2.1.1 Definition

The model we consider is as follows. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote a $d \times 1$ mean vector and a $d \times d$ covariance matrix, respectively. Also, let $\nu > 0$ be fixed and known. Assume that, conditional on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the d-dimensional random vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are iid $t_d(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$; that is, $\boldsymbol{y}_1$ has density

$$f_{t_d}(\boldsymbol{y}_1; \nu, \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \Gamma\left(\frac{d+\nu}{2}\right)}{\pi^{\frac{d}{2}} \Gamma\left(\frac{\nu}{2}\right) \nu^{\frac{d}{2}}} \left[1 + \frac{(\boldsymbol{y}_1 - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_1 - \boldsymbol{\mu})}{\nu}\right]^{-\frac{d+\nu}{2}}.$$

Let $f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the joint density of $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$; that is,

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} f_{t_d}(\boldsymbol{y}_i; \nu, \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

We take the prior on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to be the standard noninformative prior for a multivariate location-scale problem; that is, $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{d+1}{2}}$ (see, e.g., Section 8.2 of Box and Tiao, 1992). We show later in this section that (no matter what the value of ν) the resulting posterior is proper if and only if $n \geq d + 1$. Assuming that

$n \geq d+1$, the posterior density is of course

$$\begin{aligned} f(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{m(\boldsymbol{y})} \\ &\propto |\boldsymbol{\Sigma}|^{-\frac{n+d+1}{2}} \prod_{i=1}^{n} [1 + \nu^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})]^{-\frac{d+\nu}{2}}, \end{aligned} \tag{2.1}$$

where

$$m(\boldsymbol{y}) = \int_S \int_{\mathbb{R}^d} f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\mu} d\boldsymbol{\Sigma}$$

with

$$S = \left\{ \boldsymbol{z} \in \mathbb{R}^{\frac{d(d+1)}{2}} : \text{ unvech}(\boldsymbol{z}) \text{ is p.d.} \right\},$$

where unvech($\boldsymbol{z}$) is the symmetric matrix corresponding to the elements of $\boldsymbol{z}$ and "p.d." means positive definite. Our ultimate goal will be to draw samples from (2.1), but before developing methods for sampling we will establish the propriety of the posterior.

2.1.2 Propriety of the Posterior

We will prove that $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y})$ is proper if and only if $n \geq d+1$. We start with the following lemma, which seems obvious but could not be found in the literature.

Lemma 2.1. *Suppose that $\Theta \subset \mathbb{R}^p$ and let $\{f(x|\theta) : \theta \in \Theta\}$ be a family of densities indexed by θ such that the support, $\mathcal{X} = \{x : f(x|\theta) > 0\}$, does not depend on θ. Let $x_1, \ldots, x_n$ be iid from $f(\cdot|\theta)$ and suppose that $\pi(\theta)$ is an improper prior density yielding the posterior*

$$f(\theta|x_1, \ldots, x_n) \propto \left[\prod_{i=1}^{n} f(x_i|\theta) \right] \pi(\theta)\,. \tag{2.2}$$

If there exists an $n' \in \mathbb{N} := \{1, 2, \ldots\}$ such that the posterior is proper; that is, such that

$$c(x_1, \ldots, x_{n'}) = \int_\Theta \left[\prod_{i=1}^{n'} f(x_i|\theta) \right] \pi(\theta)\, d\theta < \infty$$

for all $(x_1, \ldots, x_{n'}) \in \mathcal{X}^{n'}$, then the posterior (2.2) is proper for all $n \geq n'$.

Proof. If $n > n'$ we have,

$$\int_\Theta \left[\prod_{i=1}^{n} f(x_i|\theta)\right] \pi(\theta)\, d\theta = \int_\Theta \left[\prod_{i=n'+1}^{n} f(x_i|\theta)\right] \left[\prod_{i=1}^{n'} f(x_i|\theta)\right] \pi(\theta)\, d\theta$$
$$= c(x_1, \ldots, x_{n'}) \int_\Theta \left[\prod_{i=n'+1}^{n} f(x_i|\theta)\right] \left\{ \frac{\prod_{i=1}^{n'} f(x_i|\theta)\, \pi(\theta)}{c(x_1, \ldots, x_{n'})} \right\} d\theta\,.$$

Now since the quantity in curly brackets is a proper (prior) density on θ, the integral is finite and the posterior is proper. □

Lemma 2.1 can be stated to hold for almost all $x_1, \ldots, x_n$ (not all $x_1, \ldots, x_n$), but we will omit such technicalities even though we use the lemma when it holds only outside a set of measure 0. Here is our main result about the propriety of the posterior.

Theorem 2.1. *Let λ denote Lebesgue measure in $\mathbb{R}^{dn}$. The posterior, $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y})$, is proper for λ-almost all $\boldsymbol{y}$ if and only if $n \geq d + 1$.*

Proof. The goal is to show that

$$\int_S \int_{\mathbb{R}^d} f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma} < \infty \tag{2.3}$$

for λ-almost all $\boldsymbol{y}$, if and only if $n \geq d + 1$. Instead of analyzing this integral directly, we will analyze a related integral that is more tractable. This integral involves the augmented likelihood, which is based on the so called "missing data" $\boldsymbol{q} := (q_1, \ldots, q_n)^\top \in \mathcal{Q}$, where $\mathcal{Q}$ is usually a subset of some Euclidean space (detailed in Chapter 3)

Conditional on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, let $(\boldsymbol{y}_i, q_i)$, $i = 1, \ldots, n$, be iid pairs satisfying the relations

$$\begin{aligned} \boldsymbol{y}_i|q_i, \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \mathrm{N}_d\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{q_i}\right) \\ q_i|\boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right)\,. \end{aligned} \tag{2.4}$$

Then the implied joint density of $\boldsymbol{y}$ and $\boldsymbol{q}$ is

$$f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i|q_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) f(q_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and it is easy to see that

$$\int_{\mathbb{R}_+^n} f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{q} = f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is as in Equation (2.3). Therefore,

$$\int_S \int_{\mathbb{R}^d} f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma} = \int_S \int_{\mathbb{R}^d} \int_{\mathbb{R}_+^n} f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{q}\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}\,.$$

Of course, by Fubini's Theorem, we are allowed to integrate in any order we like (notice that the integrand is nonnegative). First,

$$\begin{aligned}
&f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \frac{\left(\frac{\nu}{2}\right)^{\frac{n\nu}{2}}}{(2\pi)^{\frac{dn}{2}}\, \Gamma^n\left(\frac{\nu}{2}\right)} \left(\prod_{j=1}^{n} q_j\right)^{\frac{\nu+d-2}{2}} |\boldsymbol{\Sigma}|^{-\frac{n+d+1}{2}} e^{-\frac{\nu}{2}q_\cdot} e^{-\frac{1}{2}\sum_{j=1}^{n} q_j (\boldsymbol{y}_j - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu})}\,,
\end{aligned}$$

where $q_\cdot := \sum_{j=1}^{n} q_j$. But

$$\begin{aligned}
&\sum_{j=1}^{n} q_j (\boldsymbol{y}_j - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}) = \sum_{j=1}^{n} q_j \boldsymbol{y}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_j - 2\sum_{j=1}^{n} q_j \boldsymbol{y}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + q_\cdot \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\
&= (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \left(\frac{\boldsymbol{\Sigma}}{q_\cdot}\right)^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}}^\top \left(\frac{\boldsymbol{\Sigma}}{q_\cdot}\right)^{-1} \hat{\boldsymbol{\mu}} + \sum_{j=1}^{n} q_j \boldsymbol{y}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_j\,,
\end{aligned}$$

where $\hat{\boldsymbol{\mu}} = q_\cdot^{-1} \sum_{j=1}^{n} q_j \boldsymbol{y}_j$. Thus,

$$\begin{aligned}
&\int_{\mathbb{R}^d} f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\mu} \\
&= \frac{\left(\frac{\nu}{2}\right)^{\frac{n\nu}{2}} \left(\prod_{j=1}^{n} q_j\right)^{\frac{\nu+d-2}{2}} (q_\cdot)^{-\frac{d}{2}}\, e^{-\frac{\nu}{2}q_\cdot}}{(2\pi)^{\frac{d(n-1)}{2}}\, \Gamma^n\left(\frac{\nu}{2}\right)} |\boldsymbol{\Sigma}|^{-\frac{n+d}{2}} e^{-\frac{1}{2}\left[\sum_{j=1}^{n} q_j \boldsymbol{y}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_j - \hat{\boldsymbol{\mu}}^\top \left(\frac{\boldsymbol{\Sigma}}{q_\cdot}\right)^{-1} \hat{\boldsymbol{\mu}}\right]}.
\end{aligned} \tag{2.5}$$

Also notice that

$$\sum_{j=1}^{n} q_j \boldsymbol{y}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_j - \hat{\boldsymbol{\mu}}^\top \left(\frac{\boldsymbol{\Sigma}}{q_\cdot}\right)^{-1} \hat{\boldsymbol{\mu}} = \operatorname{tr}\left[\sum_{j=1}^{n} q_j \boldsymbol{y}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_j - q_\cdot \hat{\boldsymbol{\mu}}^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}}\right]$$
$$= \operatorname{tr}\left[\left(\sum_{j=1}^{n} q_j \boldsymbol{y}_j \boldsymbol{y}_j^\top - q_\cdot \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^\top\right) \boldsymbol{\Sigma}^{-1}\right].$$

We now assume that $n \geq d+1$ and will show that the posterior is proper for λ-almost all $\boldsymbol{y}$; that is, we will show that (2.5) is integrable with respect to both $\boldsymbol{\Sigma}$ and $\boldsymbol{q}$. To find the integral with respect to $\boldsymbol{\Sigma}$, we use the fact that the Inverse Wishart density $f_{\mathrm{IW}_d}\left(\boldsymbol{\Sigma}; n-1, \left(\sum_{j=1}^{n} q_j \boldsymbol{y}_j \boldsymbol{y}_j^\top - q_\cdot \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^\top\right)^{-1}\right)$, defined in Section 4.3, integrates to 1. That is, we have

$$\int_S \int_{\mathbb{R}^d} f(\boldsymbol{y}, \boldsymbol{q} | \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}$$
$$= \frac{\left(\frac{\nu}{2}\right)^{\frac{n\nu}{2}} \prod_{i=1}^{d} \Gamma\left(\frac{1}{2}(n-i)\right)}{\Gamma^n\left(\frac{\nu}{2}\right) \pi^{-\frac{d}{4}(d+1-2n)}} \left(\prod_{j=1}^{n} q_j\right)^{\frac{\nu+d-2}{2}} (q_\cdot)^{-\frac{d}{2}} e^{-\frac{\nu}{2} q_\cdot} \left|\sum_{j=1}^{n} q_j \boldsymbol{y}_j \boldsymbol{y}_j^\top - q_\cdot \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^\top\right|^{-\frac{n-1}{2}}.$$

Now, for $i = 1, 2, \ldots, n$, let $p_i = \frac{q_i}{q_\cdot}$. Also, let $\boldsymbol{p}_* = (p_2, p_3, \ldots, p_n)^\top$ and let $\boldsymbol{P}_*$ be a diagonal matrix with $\boldsymbol{p}_*$ on the diagonal. Finally, set

$$\boldsymbol{M}^{(1)} = \begin{pmatrix} (\boldsymbol{y}_1 - \boldsymbol{y}_2)^\top \\ \vdots \\ (\boldsymbol{y}_1 - \boldsymbol{y}_n)^\top \end{pmatrix}.$$

Now, straightforward calculations show that

$$\frac{1}{q_\cdot}\left[\sum_{j=1}^{n} q_j \boldsymbol{y}_j \boldsymbol{y}_j^\top - q_\cdot \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^\top\right] = \sum_{j=1}^{n} p_j (\boldsymbol{y}_1 - \boldsymbol{y}_j)(\boldsymbol{y}_1 - \boldsymbol{y}_j)^\top - (\boldsymbol{y}_1 - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_1 - \hat{\boldsymbol{\mu}})^\top$$
$$= \boldsymbol{M}^{(1)\top} \boldsymbol{P}_* \boldsymbol{M}^{(1)} - \boldsymbol{M}^{(1)\top} \boldsymbol{p}_* \boldsymbol{p}_*^\top \boldsymbol{M}^{(1)}$$
$$= \boldsymbol{M}^{(1)\top} \left[\boldsymbol{P}_* - \boldsymbol{p}_* \boldsymbol{p}_*^\top\right] \boldsymbol{M}^{(1)}.$$

Hence, if $n = d + 1$, we have

$$\begin{aligned}
\left| \sum_{j=1}^{n} q_j \boldsymbol{y}_j \boldsymbol{y}_j^\top - q.\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top \right| &= \left| q.\, \boldsymbol{M}^{(1)\top} \left[\boldsymbol{P}_* - \boldsymbol{p}_* \boldsymbol{p}_*^\top \right] \boldsymbol{M}^{(1)} \right| \\
&= q.^d \left| \boldsymbol{M}^{(1)} \right|^2 \left| \boldsymbol{P}_* - \boldsymbol{p}_* \boldsymbol{p}_*^\top \right| \\
&= q.^d \left| \boldsymbol{M}^{(1)} \right|^2 \left| \boldsymbol{P}_* \right| \left| \boldsymbol{I} - \boldsymbol{P}_*^{-1} \boldsymbol{p}_* \boldsymbol{p}_*^\top \right| \\
&= q.^d \left| \boldsymbol{M}^{(1)} \right|^2 \left(\prod_{j=2}^{n} p_j \right) \left| 1 - \boldsymbol{p}_*^\top \boldsymbol{P}_*^{-1} \boldsymbol{p}_* \right| \\
&= q.^d \left| \boldsymbol{M}^{(1)} \right|^2 \prod_{j=1}^{n} p_j \\
&= q.^{-1} \left| \boldsymbol{M}^{(1)} \right|^2 \prod_{j=1}^{n} q_j \,,
\end{aligned}$$

where the fourth equation follows from an elementary property of determinants that can be found in Fang and Zhang (1990, p. 3). Therefore, when $n = d + 1$ we have

$$\begin{aligned}
&\int_S \int_{\mathbb{R}^d} f(\boldsymbol{y}, \boldsymbol{q} | \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma} \\
&\quad = \frac{\prod_{i=1}^{d} \Gamma\left(\frac{1}{2}(d + 1 - i)\right)}{\left| \boldsymbol{M}^{(1)} \right|^d \pi^{\frac{d(d+1)}{4}}} \prod_{j=1}^{n} \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} q_j^{\frac{\nu}{2}-1} e^{-\frac{q_j \nu}{2}} \,,
\end{aligned} \tag{2.6}$$

which is clearly integrable in $\boldsymbol{q}$. Thus, assuming $\left| \sum_{j=1}^{n} q_j \boldsymbol{y}_j \boldsymbol{y}_j^\top - q.\boldsymbol{\mu}\boldsymbol{\mu}^\top \right| \neq 0$ (which is true for λ-almost all $\boldsymbol{y}$), it follows that the posterior is proper when $n = d + 1$, and because of Lemma 2.1 that the posterior is proper for $n \geq d + 1$.

Now, to show that the posterior is improper when $n \leq d$, denote $\boldsymbol{Q} = \sum_{j=1}^{n} q_j \boldsymbol{y}_j \boldsymbol{y}_j^\top - q.\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top$ and notice that by the transformation $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1}$ with Jacobian $J = |\boldsymbol{Z}|^{-(d+1)}$ (see; e.g., Fang and Zhang, 1990), we have

$$\int_S e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}^{-1})} |\boldsymbol{\Sigma}|^{-\frac{n+d}{2}} d\boldsymbol{\Sigma} = \int_S e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{Q}\boldsymbol{Z})} |\boldsymbol{Z}|^{\frac{n-d-2}{2}} d\boldsymbol{Z}.$$

Since $\boldsymbol{Z}$ is a p.d. matrix, by the Cholesky decomposition, it can be uniquely represented as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{X}^\top$, where $\boldsymbol{X}$ is a lower triangular matrix (l.t.m.)

with positive diagonal elements, and the Jacobian of this transformation is $J = 2^d \prod_{i=1}^d x_{ii}^{d-i+1}$. Therefore

$$\int_S e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{QZ})} |\boldsymbol{Z}|^{\frac{n-d-2}{2}} d\boldsymbol{Z} = 2^d \int_R e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{QXX}^\top)} |\boldsymbol{X}|^{n-d-2} \prod_{i=1}^d x_{ii}^{d-i+1} d\boldsymbol{X}$$

$$= 2^d \int_R e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{QXX}^\top)} \prod_{i=1}^d x_{ii}^{n-i+1} d\boldsymbol{X} = 2^d \int_R e^{-\frac{1}{2}\sum_{i=1}^d \boldsymbol{x}_i^\top \boldsymbol{Q} \boldsymbol{x}_i} \prod_{i=1}^d x_{ii}^{n-i+1} d\boldsymbol{X},$$

where

$$R = \{(r_{11}, r_{21}, r_{22}, r_{31}, \ldots, r_{dd}) \in \mathbb{R}^{\frac{d(d+1)}{2}} : r_{ii} > 0, \ \forall\, i = 1, \ldots, d\},$$

and $\boldsymbol{x}_i$ for $i = 1, \ldots, d$ are the columns of $\boldsymbol{X}$. Recall that $n \leq d$ implying that $\boldsymbol{Q}$ is non-negative definite ($\boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x} = 0$ for at least one non-zero $\boldsymbol{x} \in \mathbb{R}^d$), which means that $\boldsymbol{Q} = \boldsymbol{T}^\top \boldsymbol{T}$, where $\boldsymbol{T}$ is a l.t.m. with non-negative diagonal elements, at least one of which is zero. Let $t_{kk} = 0$ and observe that the integral

$$\int_R e^{-\frac{1}{2}\sum_{i=1}^d \boldsymbol{x}_i^\top \boldsymbol{Q} \boldsymbol{x}_i} \prod_{i=1}^d x_{ii}^{n-i+1} d\boldsymbol{X} = \int_R e^{-\frac{1}{2}\sum_{i=1}^d (\boldsymbol{T}\boldsymbol{x}_i)^\top (\boldsymbol{T}\boldsymbol{x}_i)} \prod_{i=1}^d x_{ii}^{n-i+1} d\boldsymbol{X}$$

diverges because the coefficient in front of x_{kk} in the exponential is 0; that is,

$$\int_R e^{-\frac{1}{2}\sum_{i=1}^d (\boldsymbol{T}\boldsymbol{x}_i)^\top (\boldsymbol{T}\boldsymbol{x}_i)} \prod_{i=1}^d x_{ii}^{n-i+1} d\boldsymbol{X} = c \int_{\mathbb{R}_+} e^{-\frac{1}{2} t_{kk} x_{kk}} x_{kk}^{n-k+1} dx_{kk} = \int_{\mathbb{R}_+} x_{kk}^{n-k+1} = \infty,$$

where the first equality is justified by Fubini's theorem even if $c = \infty$. □

2.2 Making Exact Draws from the Target Posterior

As we now explain, a byproduct of the proof of Theorem 2.1 is a method for making *exact* draws from the target posterior, $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y})$, in the special case where n is exactly $d + 1$.

First notice that

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y}) = \int_{\mathbb{R}_+^n} f(\boldsymbol{q}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y}) d\boldsymbol{q},$$

where

$$f(\boldsymbol{q},\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{q},\boldsymbol{\mu},\boldsymbol{\Sigma})f(\boldsymbol{q}|\boldsymbol{\mu},\boldsymbol{\Sigma})\pi(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\int_{\mathcal{Q}}\int_{S}\int_{\mathbb{R}^n} f(\boldsymbol{y}|\boldsymbol{q},\boldsymbol{\mu},\boldsymbol{\Sigma})f(\boldsymbol{q}|\boldsymbol{\mu},\boldsymbol{\Sigma})\pi(\boldsymbol{\mu},\boldsymbol{\Sigma})d\boldsymbol{\mu}d\boldsymbol{\Sigma}d\boldsymbol{q}}.$$

Thus, our problem would be solved if we could make draws from $f(\boldsymbol{q},\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{y})$ since we could then just ignore the $\boldsymbol{q}$ component. Now, since $f(\boldsymbol{q},\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{y}) = f(\boldsymbol{q}|\boldsymbol{y})f(\boldsymbol{\Sigma}|\boldsymbol{q},\boldsymbol{y})f(\boldsymbol{\mu}|\boldsymbol{\Sigma},\boldsymbol{q},\boldsymbol{y})$, we can make draws from $f(\boldsymbol{q},\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{y})$ sequentially. Actually the draws from $f(\boldsymbol{\Sigma}|\boldsymbol{q},\boldsymbol{y})$ and $f(\boldsymbol{\mu}|\boldsymbol{\Sigma},\boldsymbol{q},\boldsymbol{y})$ are trivial, since the calculations in the proof of Theorem 2.1 show that

$$\boldsymbol{\Sigma}|\boldsymbol{q},\boldsymbol{y} \sim \mathrm{IW}_d\left[n-1,\left(\sum_{j=1}^{n} q_j(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})^{\top}\right)^{-1}\right],$$

and

$$\boldsymbol{\mu}|\boldsymbol{\Sigma},\boldsymbol{q},\boldsymbol{y} \sim \mathrm{N}_d\left(\hat{\boldsymbol{\mu}},\frac{\boldsymbol{\Sigma}}{q}\right),$$

and exact methods for sampling from these distributions are well known. Furthermore, Equation (2.6) shows that when $n = d+1$, conditional on $\boldsymbol{y}$, the q_is are iid $\Gamma(\nu/2,\nu/2)$. Hence, one can make exact draws from $f(\boldsymbol{q},\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{y}) = f(\boldsymbol{q}|\boldsymbol{y})f(\boldsymbol{\Sigma}|\boldsymbol{q},\boldsymbol{y})f(\boldsymbol{\mu}|\boldsymbol{\Sigma},\boldsymbol{q},\boldsymbol{y})$ sequentially, using the following three steps:

1. Draw $q_1,\ldots,q_n$ independently from the $\Gamma(\nu/2,\nu/2)$ distribution.
2. Draw $\boldsymbol{\Sigma} \sim \mathrm{IW}_d\left[n-1,\left(\sum_{j=1}^{n} q_j(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})^{\top}\right)^{-1}\right]$.
3. Draw $\boldsymbol{\mu} \sim \mathrm{N}_d\left(\hat{\boldsymbol{\mu}},\frac{\boldsymbol{\Sigma}}{q}\right)$.

Marginally, the resulting $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is an exact draw from $f(\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{y})$. Unfortunately, when $n > d+1$, the q_is are no longer conditionally independent, and thus drawing from $f(\boldsymbol{q}|\boldsymbol{y})$ is no longer straightforward. However, as we will show, it is possible to construct rejection samplers for $f(\boldsymbol{q}|\boldsymbol{y})$ that are reasonably efficient when d is small.

We now consider the problem of sampling from

$$f(\boldsymbol{q}|\boldsymbol{y}) \propto \left(\prod_{j=1}^{n} q_j\right)^{\frac{\nu+d-2}{2}} (q.)^{-\frac{d}{2}} e^{-\frac{\nu}{2}q.} \left|\sum_{j=1}^{n} q_j \boldsymbol{y}_j \boldsymbol{y}_j^\top - q.\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top\right|^{-\frac{n-1}{2}}. \tag{2.7}$$

Notice that Equation (2.7) can also be written as

$$f(\boldsymbol{q}|\boldsymbol{y}) \propto \left(\prod_{i=1}^{n} q_i\right)^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}\sum_{i=1}^{n} q_i} \frac{\left(\prod_{i=1}^{n} q_i\right)^{\frac{d}{2}}}{q.^{\frac{d}{2}} \left|\sum_{j=1}^{n} q_j \boldsymbol{y}_j \boldsymbol{y}_j^\top - q.\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top\right|^{\frac{n-1}{2}}}. \tag{2.8}$$

But

$$\sum_{j=1}^{n} q_j \boldsymbol{y}_j \boldsymbol{y}_j^\top - q.\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top = q. \sum_{j=1}^{n} \frac{q_j}{q.} (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})^\top =: q.C(\boldsymbol{q}, \boldsymbol{y}),$$

and therefore Equation (2.8) becomes

$$\begin{aligned} f(\boldsymbol{q}|\boldsymbol{y}) &\propto \left(\prod_{i=1}^{n} q_i\right)^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}\sum_{i=1}^{n} q_i} \frac{\left(\prod_{i=1}^{n} q_i\right)^{\frac{d}{2}}}{q.^{\frac{d}{2}} |q.C(\boldsymbol{q}, \boldsymbol{y})|^{\frac{n-1}{2}}} \\ &= \left(\prod_{i=1}^{n} q_i\right)^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}\sum_{i=1}^{n} q_i} \frac{\left(\prod_{i=1}^{n} \frac{q_i}{q.}\right)^{\frac{d}{2}}}{|C(\boldsymbol{q}, \boldsymbol{y})|^{\frac{n-1}{2}}} \\ &\propto f_{\Gamma(\frac{\nu}{2}, \frac{\nu}{2})}(q_1) \dots f_{\Gamma(\frac{\nu}{2}, \frac{\nu}{2})}(q_n) \sqrt{\frac{\left(\prod_{i=1}^{n} \frac{q_i}{q.}\right)^{d}}{|C(\boldsymbol{q}, \boldsymbol{y})|^{n-1}}}. \end{aligned} \tag{2.9}$$

Since the first part is just a product of iid $\Gamma(\frac{\nu}{2}, \frac{\nu}{2})$ densities, we might consider a rejection sampler in which the candidate is exactly this product of Gamma densities. To do this, we need a bound for

$$\frac{\left(\prod_{i=1}^{n} \frac{q_i}{q.}\right)^{d}}{|C(\boldsymbol{q}, \boldsymbol{y})|^{n-1}}. \tag{2.10}$$

The following lemmas are needed to establish our bound.

Lemma 2.2. *Suppose that $\boldsymbol{Y}$ is a discrete random vector putting mass p_i at $\boldsymbol{y}_i$, $i = 1, 2, \dots, n$. Then*

$$cov(\boldsymbol{Y}) = \sum_{1 \le i < j \le n} p_i p_j (\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top.$$

Proof. Let $\boldsymbol{X}$ be an independent copy of $\boldsymbol{Y}$. Then notice that

$$\begin{aligned}
&\frac{1}{2}\mathrm{E}(\boldsymbol{X}-\boldsymbol{Y})(\boldsymbol{X}-\boldsymbol{Y})^\top = \frac{1}{2}\mathrm{E}(\boldsymbol{X}-\boldsymbol{\mu}+\boldsymbol{\mu}-\boldsymbol{Y})(\boldsymbol{X}-\boldsymbol{\mu}+\boldsymbol{\mu}-\boldsymbol{Y})^\top \\
&= \frac{1}{2}\mathrm{E}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top + \frac{1}{2}\mathrm{E}(\boldsymbol{\mu}-\boldsymbol{Y})(\boldsymbol{\mu}-\boldsymbol{Y})^\top \\
&+ \frac{1}{2}\mathrm{E}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{\mu}-\boldsymbol{Y})^\top + \frac{1}{2}\mathrm{E}(\boldsymbol{\mu}-\boldsymbol{Y})(\boldsymbol{X}-\boldsymbol{\mu})^\top \\
&= \frac{1}{2}\mathrm{cov}(\boldsymbol{X}) + \frac{1}{2}\mathrm{cov}(\boldsymbol{Y}) = \mathrm{cov}(\boldsymbol{Y}),
\end{aligned}$$

where $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Y}) = \mathrm{E}(\boldsymbol{X})$. Now apply Equation (2.2) to get

$$\begin{aligned}
\mathrm{cov}(\boldsymbol{Y}) &= \frac{1}{2}\mathrm{E}(\boldsymbol{X}-\boldsymbol{Y})(\boldsymbol{X}-\boldsymbol{Y})^\top = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (\boldsymbol{y}_i-\boldsymbol{y}_j)(\boldsymbol{y}_i-\boldsymbol{y}_j)^\top \\
&= \sum_{i<j} p_i p_j (\boldsymbol{y}_i-\boldsymbol{y}_j)(\boldsymbol{y}_i-\boldsymbol{y}_j)^\top.
\end{aligned}$$

□

Before we proceed any further, we need to develop some notation. For $m \leq n$, let C_m^n be the set of all possible subsets of size m from $\mathbb{N}_n := \{1, 2, \ldots, n\}$; that is,

$$C_m^n := \{(k_1, \ldots, k_m) : k_1 < \cdots < k_m \text{ and } k_i \in \mathbb{N}_n, i = 1, \ldots, m\}.$$

For example, $C_2^3 := \{(1,2), (1,3), (2,3)\}$. Also, for a given matrix $\boldsymbol{A}$ of dimension $m \times n$, let $\boldsymbol{A}_{r(T)}$ denote the submatrix obtained from $\boldsymbol{A}$ using only the rows that are in the index set T (of course, here $T \subset \mathbb{N}_m$). Similarly, let $\boldsymbol{A}_{c(T)}$ be the submatrix obtained from the columns of $\boldsymbol{A}$ that are in the set T. We now state the Binet-Cauchy formula for the determinant of a product of two matrices. It can be found, for example, in Karlin (1968).

Lemma 2.3 (Binet-Cauchy Formula). *Let $\boldsymbol{A}$ be an $m \times n$ matrix, and let $\boldsymbol{B}$ be an $n \times m$ matrix. Then*

$$|\boldsymbol{AB}| = \sum_{T \in C_m^n} |\boldsymbol{A}_{c(T)}| \, |\boldsymbol{B}_{r(T)}|.$$

Let's now go back to the question of how to bound the expression in Equation (2.10).

Theorem 2.2. *For $d \in \mathbb{N}$, $n \geq d+1$, $\boldsymbol{y} \in \mathbb{R}^{dn}$ and $\boldsymbol{q} \in \mathbb{R}^n_+$,*

$$\frac{\left(\prod_{i=1}^n \frac{q_i}{q_\cdot}\right)^d}{|C(\boldsymbol{q},\boldsymbol{y})|^{n-1}} \leq \frac{1}{\binom{n-1}{d}^{n-1} d(\boldsymbol{y})}, \tag{2.11}$$

where

$$d(\boldsymbol{y}) = \begin{cases} \frac{1}{(y_1-y_2)^2} & \text{if } n=2 \\ \left(\frac{1}{c_1(\boldsymbol{y})^{\frac{1}{d(n-2)}}} + \cdots + \frac{1}{c_n(\boldsymbol{y})^{\frac{1}{d(n-2)}}}\right)^{-d(n-2)} & \text{if } n>2, \end{cases}$$

$$c_i(\boldsymbol{y}) = \prod_{T \in C_d^{n-1}} \left|\boldsymbol{M}^{(i)}_{r(T)}\right|^{\frac{2d!(n-d-1)!}{(n-2)!}},$$

and $\boldsymbol{M}^{(i)}$ is an $(n-1)\times d$ matrix, given by

$$\boldsymbol{M}^{(i)} = \begin{pmatrix} (\boldsymbol{y}_i - \boldsymbol{y}_1)^\top \\ \vdots \\ (\boldsymbol{y}_i - \boldsymbol{y}_{i-1})^\top \\ (\boldsymbol{y}_i - \boldsymbol{y}_{i+1})^\top \\ \vdots \\ (\boldsymbol{y}_i - \boldsymbol{y}_n)^\top \end{pmatrix},$$

for $i = 1, \ldots, n$.

Proof. First notice that when $n = 2$, d must be 1, and in this case both sides of (2.11) are equal to $\frac{1}{(y_1-y_2)^2}$. For the nontrivial case when $n > 2$, denote $p_i = \frac{q_i}{q_\cdot}$, $i = 1, \ldots, n$ and notice that because of Lemma 2.2

$$\sup_{\boldsymbol{q}\in\mathbb{R}^n_+} \frac{\left(\prod_{i=1}^n \frac{q_i}{q_\cdot}\right)^d}{|C(\boldsymbol{q},\boldsymbol{y})|^{n-1}} = \sup_{\boldsymbol{p}\in\mathcal{P}} \frac{\left(\prod_{i=1}^n p_i\right)^d}{\left|\sum_{i<j} p_i p_j (\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top\right|^{n-1}},$$

where $\mathcal{P} = \{(p_1, \ldots, p_n) \in (0,1)^n : \sum_{i=1}^n p_i = 1\}$.

What follows is true for any $p_k, k = 1, \ldots, n$; but for ease of exposition, we'll show the details only for p_1. First, notice that in the sense of matrix ordering[1],

$$\begin{aligned}\sum_{i<j} p_i p_j (\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top &\geq \sum_{i=2}^{n} p_1 p_i (\boldsymbol{y}_1 - \boldsymbol{y}_i)(\boldsymbol{y}_1 - \boldsymbol{y}_i)^\top \\ &= p_1 \boldsymbol{M}^{(1)^\top} \boldsymbol{P}_* \boldsymbol{M}^{(1)},\end{aligned} \tag{2.12}$$

where $\boldsymbol{P}_*$ is as defined in Section 2.1. Now applying the Binet-Cauchy formula with $\boldsymbol{A} = \boldsymbol{M}^{(1)^\top}$ and $\boldsymbol{B} = \boldsymbol{P}_* \boldsymbol{M}^{(1)}$, we obtain

$$\left| \boldsymbol{M}^{(1)^\top} \boldsymbol{P}_* \boldsymbol{M}^{(1)} \right| = \sum_{T \in C_d^{n-1}} \left| \left(\boldsymbol{M}^{(1)^\top} \right)_{c(T)} \right| \left| \left(\boldsymbol{P}_* \boldsymbol{M}^{(1)} \right)_{r(T)} \right|. \tag{2.13}$$

Since

$$\boldsymbol{P}_* \boldsymbol{M}^{(1)} = \begin{pmatrix} p_2 (\boldsymbol{y}_1 - \boldsymbol{y}_2)^\top \\ \cdots \\ p_n (\boldsymbol{y}_1 - \boldsymbol{y}_n)^\top \end{pmatrix},$$

for any $T \in C_d^{n-1}$, we can write

$$\left| \left(\boldsymbol{P}_* \boldsymbol{M}^{(1)} \right)_{r(T)} \right| = \left(\prod_{i \in T} p_{i+1} \right) \left| \boldsymbol{M}^{(1)}_{r(T)} \right|.$$

Thus,

$$\left| \boldsymbol{M}^{(1)^\top} \boldsymbol{P}_* \boldsymbol{M}^{(1)} \right| = \sum_{T \in C_d^{n-1}} \left(\prod_{i \in T} p_{i+1} \right) \left| \boldsymbol{M}^{(1)}_{r(T)} \right|^2,$$

and consequently from Equations (2.12), (2.13) and the fact that if $\boldsymbol{A} \geq \boldsymbol{B}$ then $|\boldsymbol{A}| \geq |\boldsymbol{B}|$, it follows that

$$\left| \sum_{i<j} p_i p_j (\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top \right| \geq p_1^d \sum_{T \in C_d^{n-1}} \left(\prod_{i \in T} p_{i+1} \right) \left| \boldsymbol{M}^{(1)}_{r(T)} \right|^2. \tag{2.14}$$

[1] For nonnegative definite $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \geq \boldsymbol{B}$ iff $\boldsymbol{A} - \boldsymbol{B}$ is nonnegative definite

Now by the inequality between the arithmetic and geometric means, we obtain

$$\begin{aligned}&\sum_{T\in C_d^{n-1}}\left(\prod_{i\in T}p_{i+1}\right)\left|M^{(1)}_{r(T)}\right|^2\\ &\qquad\geq\binom{n-1}{d}\left[\prod_{T\in C_d^{n-1}}\left(\prod_{i\in T}p_{i+1}\right)\left|M^{(1)}_{r(T)}\right|^2\right]^{\binom{n-1}{d}^{-1}}.\end{aligned}\tag{2.15}$$

Notice that each index $i\in\{1,\dots,n\}$ is contained in exactly $\binom{n-2}{d-1}$ elements of the set C_d^{n-1}. Therefore,

$$\prod_{T\in C_d^{n-1}}\prod_{i\in T}p_{i+1}=(p_2\dots p_n)^{\binom{n-2}{d-1}},$$

implying that

$$\begin{aligned}&\left[\prod_{T\in C_d^{n-1}}\left(\prod_{i\in T}p_{i+1}\right)\left|M^{(1)}_{r(T)}\right|^2\right]^{\binom{n-1}{d}^{-1}}\\ &\qquad=(p_2\dots p_n)^{\frac{d}{n-1}}\left(\prod_{T\in C_d^{n-1}}\left|M^{(1)}_{r(T)}\right|^2\right)^{\binom{n-1}{d}^{-1}}.\end{aligned}\tag{2.16}$$

From Equations (2.14), (2.15) and (2.16) we obtain that

$$\left|\sum_{i<j}p_ip_j(\boldsymbol{y}_i-\boldsymbol{y}_j)(\boldsymbol{y}_i-\boldsymbol{y}_j)^\top\right|^{n-1}\geq\binom{n-1}{d}^{n-1}p_1^{d(n-1)}(p_2\dots p_n)^dc_1(\boldsymbol{y}),$$

and consequently

$$\frac{\left(\prod_{i=1}^n p_i\right)^d}{\left|\sum_{i<j}p_ip_j(\boldsymbol{y}_i-\boldsymbol{y}_j)(\boldsymbol{y}_i-\boldsymbol{y}_j)^\top\right|^{n-1}}\leq\frac{1}{\binom{n-1}{d}^{n-1}p_1^{d(n-2)}c_1(\boldsymbol{y})}.$$

By symmetry, for any $k\in\mathbb{N}_n$

$$\frac{\left(\prod_{i=1}^n p_i\right)^d}{\left[\sum_{i<j}p_ip_j(\boldsymbol{y}_i-\boldsymbol{y}_j)(\boldsymbol{y}_i-\boldsymbol{y}_j)^\top\right]^{n-1}}\leq\frac{1}{\binom{n-1}{d}^{n-1}p_k^{d(n-2)}c_k(\boldsymbol{y})}.$$

Thus,

$$\frac{\left(\prod_{i=1}^{n} p_i\right)^d}{\left[\sum_{i<j} p_i p_j (\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top\right]^{n-1}} \leq \min_{1\leq k\leq n} \frac{1}{\binom{n-1}{d}^{n-1} p_k^{d(n-2)} c_k(\boldsymbol{y})},$$

and therefore

$$\sup_{\boldsymbol{q}\in\mathbb{R}_+^n} \frac{\left(\prod_{i=1}^{n} \frac{q_i}{q_\cdot}\right)^d}{|C(\boldsymbol{q},\boldsymbol{y})|^{n-1}} \leq \sup_{\boldsymbol{p}\in\mathcal{P}} \min_{1\leq k\leq n} \frac{1}{\binom{n-1}{d}^{n-1} p_k^{d(n-2)} c_k(\boldsymbol{y})}.$$

Finally, we will show that

$$\sup_{\boldsymbol{p}\in\mathcal{P}} \min_{1\leq i\leq n} \frac{1}{p_i^{d(n-2)} c_i(\boldsymbol{y})} = \frac{1}{d(\boldsymbol{y})}. \tag{2.17}$$

Let

$$\begin{aligned} b_i &= c_i(\boldsymbol{y})^{-\frac{1}{d(n-2)}} \\ b_\cdot &= \sum_{i=1}^{n} b_i \\ p_i^* &= \frac{b_i}{b_\cdot} \end{aligned}$$

Then

$$\begin{aligned} \sup_{\boldsymbol{p}\in\mathcal{P}} \min_{1\leq i\leq n} \frac{1}{p_i^{d(n-2)} c_i(\boldsymbol{y})} &= \left[\inf_{\boldsymbol{p}\in\mathcal{P}} \max_{1\leq i\leq n} p_i c_i(\boldsymbol{y})^{\frac{1}{d(n-2)}}\right]^{-d(n-2)} \\ &= b_\cdot^{d(n-2)} \left(\inf_{\boldsymbol{p}\in\mathcal{P}} \max_{1\leq i\leq n} \frac{p_i}{p_i^*}\right)^{-d(n-2)}. \end{aligned}$$

Assuming $p_i^* > 0 \;\forall\, i \in \mathbb{N}_n$, it is clear that if $\boldsymbol{p} = \boldsymbol{p}^*$, then

$$\max_{1\leq i\leq n} \frac{p_i}{p_i^*} = 1,$$

whereas for any other choice of $\boldsymbol{p}$,

$$\max_{1\leq i\leq n} \frac{p_i}{p_i^*} < 1,$$

since if $p_i \neq p_i^*$ for some i, then there exists a $j \in \mathbb{N}_n$ such that $p_j > p_j^*$. □

In the univariate case ($d = 1$), the constants $c_i(\boldsymbol{y})$ can be simplified considerably. We state this simplification of Theorem 2.2 as a corollary.

Corollary 2.1. *For $n \geq 2$, $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{q} \in \mathbb{R}^n_+$,*

$$\frac{\prod_{i=1}^n \frac{q_i}{q_\cdot}}{[v(\boldsymbol{q},\boldsymbol{y})]^{n-1}} \leq \left[(n-1)^{n-1} d_1(\boldsymbol{y})\right]^{-1},$$

where

$$v(\boldsymbol{q},\boldsymbol{y}) = \frac{1}{q_\cdot}\sum_{i=1}^n q_i(y_i - \hat{\mu})^2$$

$$d_1(\boldsymbol{y}) = \left[\frac{1}{e_1(\boldsymbol{y})^{\frac{1}{n-2}}} + \cdots + \frac{1}{e_n(\boldsymbol{y})^{\frac{1}{n-2}}}\right]^{-(n-2)},$$

and

$$e_i(\boldsymbol{y}) = \prod_{j\neq i}(y_i - y_j)^2,\ i = 1,\ldots,n.$$

Proof. First notice that $C_1^{n-1} := \{1,\ldots,n-1\}$ and therefore

$$\prod_{j\in C_1^{n-1}} \left|M^{(i)}_{r(j)}\right| = \prod_{j\neq i}(y_i - y_j)^2.$$

The rest of the proof is just substitution. □

The bound proved in Theorem 2.2 is not sharp, which reduces the acceptance rate of the resulting rejection sampler. When $d = 1$, we were able to find the exact supremum when $n \in \{2,3\}$ and this result is stated in Proposition 2.1. Unfortunately, we were unable to get a better bound in the general case, but we can state a conjecture about the exact supremum of (2.10) when $d = 1$, which is presented in Conjecture 2.1

Proposition 2.1. *For $n \in \{2,3\}, \boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{q} \in \mathbb{R}^n_+$,*

$$\sup_{\boldsymbol{q}\in\mathbb{R}^n_+} \frac{\prod_{i=1}^n \left(\frac{q_i}{q_\cdot}\right)}{[v(\boldsymbol{q},\boldsymbol{y})]^{n-1}} = \left[(n-1)^{n-1} c(\boldsymbol{y})\right]^{-1}, \tag{2.18}$$

where

$$c(\boldsymbol{y}) = \min_{1\le i\le n} \prod_{j\neq i} (y_i - y_j)^2. \tag{2.19}$$

Proof. Notice that when $n = 2, v(\boldsymbol{q}, \boldsymbol{y}) = p_1p_2(y_1 - y_2)^2, c(\boldsymbol{y}) = (y_1 - y_2)^2$ and consequently (2.18) is obvious. The non-trivial case is when $n = 3$.

We will first prove that

$$\frac{\prod_{i=1}^n \binom{q_i}{q}}{[v(\boldsymbol{q}, \boldsymbol{y})]^{n-1}} \le \left[(n-1)^{n-1} c(\boldsymbol{y})\right]^{-1} \tag{2.20}$$

and then we will establish the attainment of the sup. Suppose that

$$\min_{1\le i\le 3} \prod_{j\neq i} (y_i - y_j)^2 = \prod_{j\neq 1} (y_1 - y_j)^2;$$

i.e., the minimum is achieved at $i = 1$. This simply means that y_1 is between y_2 and y_3. Without loss of generality, we will assume that the ordering of y_is is $y_2 \le y_1 \le y_3$. Notice that (2.20) is now equivalent to

$$\sqrt{\frac{p_1p_2}{p_3}}(y_1 - y_2)^2 + \sqrt{\frac{p_1p_3}{p_2}}(y_1 - y_3)^2 + \sqrt{\frac{p_2p_3}{p_1}}(y_2 - y_3)^2 \ge 2(y_1 - y_2)(y_3 - y_1). \tag{2.21}$$

But

$$(y_2 - y_3)^2 = (y_1 - y_2)^2 + (y_3 - y_1)^2 + 2(y_3 - y_1)(y_1 - y_2),$$

and because of the assumed ordering, (2.21) is equivalent to

$$\frac{1}{2}\frac{y_1 - y_2}{y_3 - y_1}\left(\sqrt{\frac{p_1p_2}{p_3}} + \sqrt{\frac{p_2p_3}{p_1}}\right) + \frac{1}{2}\frac{y_3 - y_1}{y_1 - y_2}\left(\sqrt{\frac{p_1p_3}{p_2}} + \sqrt{\frac{p_2p_3}{p_1}}\right) + \sqrt{\frac{p_2p_3}{p_1}} \ge 1.$$

Now an application of the inequality between the arithmetic and geometric mean gives us

$$\begin{aligned}&\frac{1}{2}\frac{y_1 - y_2}{y_3 - y_1}\left(\sqrt{\frac{p_1p_2}{p_3}} + \sqrt{\frac{p_2p_3}{p_1}}\right) + \frac{1}{2}\frac{y_3 - y_1}{y_1 - y_2}\left(\sqrt{\frac{p_1p_3}{p_2}} + \sqrt{\frac{p_2p_3}{p_1}}\right)\\ &\qquad \ge \sqrt{\left(\sqrt{\frac{p_1p_2}{p_3}} + \sqrt{\frac{p_2p_3}{p_1}}\right)\left(\sqrt{\frac{p_1p_3}{p_2}} + \sqrt{\frac{p_2p_3}{p_1}}\right)} = \sqrt{1 + \frac{p_2p_3}{p_1}}.\end{aligned}$$

Consequently, (2.20) will be established, if we can show that

$$\sqrt{1+\frac{p_2p_3}{p_1}}+\sqrt{\frac{p_2p_3}{p_1}}\geq 1,$$

which is obviously true.

Notice that the last inequality becomes equality when $p_1 \to 1$ and $p_{2,3} \to 0$ with the rest of the inequalities suggesting that $\frac{p_2}{p_3} = \frac{(y_3-y_1)^2}{(y_1-y_2)^2}$. It is easy to verify that in terms of q_is this is satisfied when, for example, we take

$$q_1 = t,\ q_2 = \frac{1}{t} \text{ and } q_3 = \frac{(y_1-y_2)^2}{t(y_3-y_1)^2}$$

and let $t \to \infty$. Then

$$\frac{q_1}{q_\cdot} = \frac{t}{t+\frac{1}{t}+\frac{(y_1-y_2)^2}{t(y_3-y_1)^2}} \to 1$$

$$\frac{q_2q_3}{q_1q_\cdot} = \frac{\frac{(y_1-y_2)^2}{t^2(y_3-y_1)^2}}{t\left[t+\frac{1}{t}+\frac{(y_1-y_2)^2}{t(y_3-y_1)^2}\right]} \to 0$$

as $t \to \infty$, implying that

$$\lim_{t\to\infty}\frac{1}{2}\frac{y_1-y_2}{y_3-y_1}\left(\sqrt{\frac{p_1p_2}{p_3}}+\sqrt{\frac{p_2p_3}{p_1}}\right)+\frac{1}{2}\frac{y_3-y_1}{y_1-y_2}\left(\sqrt{\frac{p_1p_3}{p_2}}+\sqrt{\frac{p_2p_3}{p_1}}\right)+\sqrt{\frac{p_2p_3}{p_1}}=1.$$

□

Conjecture 2.1. *For $n \geq 2, \boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{q} \in \mathbb{R}^n_+$,*

$$\sup_{\boldsymbol{q}\in\mathbb{R}^n_+}\frac{\prod_{i=1}^n\left(\frac{q_i}{q_\cdot}\right)}{[v(\boldsymbol{q},\boldsymbol{y})]^{n-1}} = \left[(n-1)^{n-1}c(\boldsymbol{y})\right]^{-1}, \tag{2.22}$$

where

$$c(\boldsymbol{y}) = \min_{1\leq i\leq n}\prod_{j\neq i}(y_i-y_j)^2. \tag{2.23}$$

Remark 2.1. *Notice that the result of Theorem 2.2 provides an alternative proof of the fact that the posterior is proper for $n \geq d+1$, because it shows that (2.6) is bounded from above by a quantity integrable in $\boldsymbol{q}$.*

2.3 Simulation Examples

The purpose of this Section is to study the acceptance rate of the rejection sampler proposed in Section 2.2 as a method for obtaining draws from our target density. Clearly, the bounds established in Lemma 2.2 and the subsequent corollary and conjecture depend on $\boldsymbol{y}, n$ and d. Furthermore, the candidate density depends on ν. It is, therefore, interesting to investigate the behavior of the algorithm when changing the values of these parameters. Generally speaking, the results of our computer simulations show that the acceptance rate of the algorithm varies greatly depending on the specific data set, $\boldsymbol{y}$, and suggest that it works better if either n is small and/or ν is big. Also, the acceptance rate tends to be higher when the y_is are equally spaced. Here are some specific examples.

Example 2.1: Suppose a particular set of y_is is given by[2]

$$\boldsymbol{y} = (-1.449605, -0.996631, 0.228872, 0.068414, -0.126978, -0.563358, 0.766889)^\top$$

and is assumed to have t_1 distribution with $\nu = 5$. A computer simulation using the bound in the conjecture resulted in an Accept - Reject algorithm with acceptance rate of 0.0384 estimated with a standard error of 0.00061 (using the well-known formula for the standard deviation of one observation from the Binomial distribution). Using the bound from Lemma 2.2 on the same data, results in acceptance rate of 0.00131 estimated with a standard error of 0.00011. Assuming that $\nu = 10$, changes the acceptance rates for the unproved and proved bound to 0.04141 (0.00063) and 0.00139 (0.00012), respectively (the numbers in parenthesis are the standard errors).

Example 2.2: For this example we simulated Student's t data with parameters $\nu = 3, d = 2$, and $n = 7$:

---

[2] The actual values were obtained by simulation

$\boldsymbol{y}_1 = (-1.0082, 1.5155)^\top$, $\boldsymbol{y}_2 = (-0.39561, 0.066761)^\top$

$\boldsymbol{y}_3 = (1.4634, 2.5265)^\top$, $\boldsymbol{y}_4 = (1.5163, 1.5953)^\top$

$\boldsymbol{y}_5 = (-1.0334, 3.0844)^\top$, $\boldsymbol{y}_6 = (-1.5207, 0.21293)^\top$

$\boldsymbol{y}_7 = (6.0723, -0.069308)^\top$,

which produced a bound of 0.000668735. Table 2.3 gives the acceptance rates of the rejection sampler for different values of ν along with their estimated standard errors and is based on 1,000,000 iterations.

Table 2–1: Acceptance rates of the rejection sampler when $d = 2$

ν	Acceptance rate	Standard Error
1	1.1×10^{-6}	3.31662×10^{-7}
5	4.1×10^{-6}	6.40311×10^{-7}
10	4.1×10^{-6}	6.40311×10^{-7}
15	5.1×10^{-6}	7.14141×10^{-7}

Example 2.3: In our final example, we first obtained another simulated sample from the multivariate Student's t distribution, but this time with $\nu = 5, d = 3$, and $n = 6$:

$\boldsymbol{y}_1 = (-0.92497, -0.63609, -0.058751)^\top$, $\boldsymbol{y}_2 = (0.32547, 0.0031071, 0.98505)^\top$

$\boldsymbol{y}_3 = (0.77057, 1.8546, -0.086936)^\top$, $\boldsymbol{y}_4 = (1.5351, -1.1889, -1.6686)^\top$

$\boldsymbol{y}_5 = (0.037699, -1.1527, 0.29751)^\top$, $\boldsymbol{y}_6 = (0.14760, 0.51222, -1.9081)^\top$,

which produced a bound of 0.0465003.

Here are the acceptance rates for different values of ν, based on 10,000,000 iterations along with their standard errors:

Table 2–2: Acceptance rates of the rejection sampler when $d = 3$

ν	Acceptance rate	Standard Error
1	2.3×10^{-6}	4.79583×10^{-7}
5	6.8×10^{-6}	8.24618×10^{-7}
10	6.7×10^{-6}	8.18533×10^{-7}
15	8.3×10^{-6}	9.1104×10^{-7}

CHAPTER 3
DATA-AUGMENTATION ALGORITHMS

3.1 Introduction

In two related papers (Meng and van Dyk, 1999; van Dyk and Meng, 2001), X.-L. Meng and D.A. van Dyk developed two generalizations of the standard data augmentation (DA) algorithm. The basic idea underlying these generalizations is to construct a large class of candidate augmentation schemes by introducing a *working parameter*, α, into the standard augmentation scheme. In *conditional augmentation*, a particular value of α is chosen, while in *marginal augmentation*, α is actually incorporated into the augmented model by way of a *working prior*, $\omega(\alpha)$. In all of the examples considered by Meng and van Dyk, the marginal augmentation (MA) algorithm becomes more efficient as the working prior approaches a special *improper* form. The basic MCMC theory breaks down, however, when the working prior is improper. Indeed, when $\omega(\alpha)$ is improper, the Markov chain driving the MA algorithm is *not positive recurrent*. On the other hand, this non-positive recurrent chain is on an augmented state space that is larger than the parameter space. It turns out that, under regularity conditions, the non-positive recurrent chain possesses a well behaved marginal chain that has the correct (proper) invariant distribution. In this chapter, we consider in detail the general MA algorithm and show how it can be applied to the Student's t problem. The chapter is organized as follows.

In the next section we describe the standard data augmentation algorithm for dealing with intractable posterior densities like (2.1). Then in Section 3.3 we describe the generalizations of Meng and van Dyk. One drawback of their methods is that it is very hard to verify the conditions that guarantee convergence of the

algorithm to the target density. In Section 3.4 we provide an alternative validation of the MA method that is inspired by the work of Liu and Wu (1999). Finally, in Sections 3.5 and 3.6 we highlight some problems with the proofs in van Dyk and Meng (2001). Throughout the chapter we present all the algorithms in their full generality and we use the Student's t model as an example.

3.2 Standard Data Augmentation

Suppose that $f(\boldsymbol{y}|\boldsymbol{\theta})$ is a likelihood and $\pi(\boldsymbol{\theta})$ is a prior, with supports $\mathcal{Y}$ and Θ, respectively, yielding the proper posterior density

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{m(\boldsymbol{y})}, \tag{3.1}$$

where $m(\boldsymbol{y}) := \int_\Theta f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} < \infty$. Unless $\pi(\cdot)$ is a conjugate prior, the denominator, $m(\boldsymbol{y})$, is typically an intractable integral and hence expectations with respect to this posterior (required for Bayesian inference) are not available in closed form. The posterior (2.1) is an example. In such cases, one way to analyze the posterior distribution is through Monte Carlo methods. Assuming that direct (iid) sampling is not an option, one might resort to an MCMC method where dependent samples are generated using a Markov chain with stationary density (3.1), and these samples are used to approximate expectations with respect to $f(\boldsymbol{\theta}|\boldsymbol{y})$. A good overview of MCMC methods can be found in Tierney (1994).

The MCMC algorithms developed by Meng and van Dyk are extensions of the standard DA algorithm so this is where we begin. The basic idea of DA is to construct *unobserved data*, $\boldsymbol{q} \in \mathcal{Q}$, that serve as a computational tool. To be specific, we introduce a joint density $f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\theta})$, with support $\mathcal{Y} \times \mathcal{Q}$, such that (i) the marginal density of $\boldsymbol{y}$ implied by $f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\theta})$ coincides with the original likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$; that is,

$$\int_{\mathcal{Q}} f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\theta})\,d\boldsymbol{q} = f(\boldsymbol{y}|\boldsymbol{\theta})\ ,\forall\ \boldsymbol{\theta} \in \Theta, \forall\ \boldsymbol{y} \in \mathcal{Y}, \tag{3.2}$$

and (ii) sampling from both conditional densities

$$f(\boldsymbol{q}|\boldsymbol{\theta},\boldsymbol{y}) := \frac{f(\boldsymbol{y},\boldsymbol{q}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\int_{\mathcal{Q}} f(\boldsymbol{y},\boldsymbol{q}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{q}} \quad \text{and} \quad f(\boldsymbol{\theta}|\boldsymbol{q},\boldsymbol{y}) := \frac{f(\boldsymbol{y},\boldsymbol{q}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\int_{\Theta} f(\boldsymbol{y},\boldsymbol{q}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}} \tag{3.3}$$

is feasible, if not convenient. Since $f(\boldsymbol{y},\boldsymbol{q}|\boldsymbol{\theta})$ is introduced purely for computational purposes it should not alter the posterior model (3.1). To see that condition (3.2) is sufficient to guarantee this, notice that the "augmented posterior" satisfies

$$f(\boldsymbol{\theta},\boldsymbol{q}|\boldsymbol{y}) := \frac{f(\boldsymbol{y},\boldsymbol{q}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\int_{\Theta}\int_{\mathcal{Q}} f(\boldsymbol{y},\boldsymbol{q}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{q}\,d\boldsymbol{\theta}} = \frac{f(\boldsymbol{y},\boldsymbol{q}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}} = \frac{f(\boldsymbol{y},\boldsymbol{q}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\boldsymbol{y})}\,,$$

and therefore

$$\int_{\mathcal{Q}} f(\boldsymbol{\theta},\boldsymbol{q}|\boldsymbol{y})\,d\boldsymbol{q} = \frac{\int_{\mathcal{Q}} f(\boldsymbol{y},\boldsymbol{q}|\boldsymbol{\theta})d\boldsymbol{q}\;\pi(\boldsymbol{\theta})}{m(\boldsymbol{y})} = f(\boldsymbol{\theta}|\boldsymbol{y})\,, \tag{3.4}$$

our target density.

The DA algorithm is a method for obtaining a sample $\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n\}$ with the property that the sampling distribution of $\boldsymbol{\theta}_n$ converges to (3.1), in total variation distance, as $n\to\infty$. It is based on a discrete time Markov chain,

$$\Phi_{\mathrm{DA}} = \{\boldsymbol{\theta}_i\}_{i=0}^{\infty},$$

with Markov transition density

$$K_{\mathrm{DA}}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int_{\mathcal{Q}} f(\boldsymbol{\theta}|\boldsymbol{q},\boldsymbol{y}) f(\boldsymbol{q}|\boldsymbol{\theta}',\boldsymbol{y})\,d\boldsymbol{q}\,; \tag{3.5}$$

that is, given that the current state of the chain is $\boldsymbol{\theta}'$, the conditional density of the next state, $\boldsymbol{\theta}$, is (3.5). Since

$$\int_{\Theta} K_{\mathrm{DA}}(\boldsymbol{\theta}|\boldsymbol{\theta}') f(\boldsymbol{\theta}'|\boldsymbol{y}) d\boldsymbol{\theta}' = \int_{\mathcal{Q}} f(\boldsymbol{\theta}|\boldsymbol{q},\boldsymbol{y}) \left[\int_{\Theta} f(\boldsymbol{q}|\boldsymbol{\theta}',\boldsymbol{y}) f(\boldsymbol{\theta}'|\boldsymbol{y}) d\boldsymbol{\theta}'\right] d\boldsymbol{q} =$$
$$\int_{\mathcal{Q}} f(\boldsymbol{\theta}|\boldsymbol{q},\boldsymbol{y}) f(\boldsymbol{q}|\boldsymbol{y}) d\boldsymbol{q} = \int_{\mathcal{Q}} f(\boldsymbol{\theta},\boldsymbol{q}|\boldsymbol{y}) d\boldsymbol{q} = f(\boldsymbol{\theta}|\boldsymbol{y}),$$

our target density, $f(\boldsymbol{\theta}|\boldsymbol{y})$, is an invariant density for Φ_{DA}. Therefore[1], we can use Φ_{DA} to get approximate draws from $f(\boldsymbol{\theta}|\boldsymbol{y})$. More details about data augmentation can be found in Tanner and Wong (1987), Liu, Wong and Kong (1995) and Liu, Wong and Kong (1994).

There are many ways to construct a joint density that satisfies (3.2), but here we will specifically consider the following hierarchical description:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{q},\boldsymbol{\theta} &\sim f(\boldsymbol{y}|\boldsymbol{q},\boldsymbol{\theta}) \\ \boldsymbol{q}|\boldsymbol{\theta} &\sim f(\boldsymbol{q}|\boldsymbol{\theta}). \end{aligned} \tag{3.6}$$

Typically (3.6) represents some well known decomposition as is demonstrated in the next example.

Example 3.1: Consider the univariate version of the Student's t problem, described at the beginning of Section 2.1, with $\boldsymbol{\theta} = (\mu, \sigma)$; that is, assume $y_1, \ldots, y_n$ are iid $t_1(\nu, \mu, \sigma)$ with joint density $f(\boldsymbol{y}|\mu, \sigma)$. The prior is $\pi(\mu, \sigma) \propto 1/\sigma$ and the goal is to sample from the posterior $f(\mu, \sigma|\boldsymbol{y})$.

There is a simple data augmentation algorithm for this problem that is based on the well known decomposition (that we express loosely as) $y \stackrel{d}{=} \mu + \sqrt{\sigma} z/\sqrt{q}$, where $z \sim \mathrm{N}(0,1)$ and $q \sim \chi^2_\nu/\nu$ with z and q independent. To be specific, "missing data" $\boldsymbol{q} = (q_1, \ldots, q_n)^\top$ are introduced such that, conditional on (μ, σ), (y_j, q_j) are iid pairs satisfying

$$\begin{aligned} y_j|q_j, \mu, \sigma &\sim \mathrm{N}\left(\mu, \frac{\sigma}{q_j}\right) \\ q_j|\mu, \sigma &\sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \end{aligned} \tag{3.7}$$

[1] Throughout this chapter, we assume (unless otherwise stated) that all Markov chains satisfy the usual regularity conditions described in Section 4.2

where by $X \sim \Gamma(a,b)$ we mean that the density of the random variable X is proportional to $x^{a-1}e^{-bx}I_{(0,\infty)}(x)$ and $\chi^2_k \equiv \Gamma\left(\frac{k}{2},\frac{1}{2}\right)$. The reader is invited to verify that the marginal density of y_j given (μ,σ) under (3.7) is $f_{t_1}(y_j;\nu,\mu,\sigma)$ and hence, because of the independence assumption, (3.2) is satisfied. Thus, we are now in a position to apply the DA algorithm; i.e., we consider simulation of the Markov chain defined by (3.5). It is trivial to verify that, conditional on $(\mu,\sigma,\boldsymbol{y})$, the random variables $q_1,\ldots,q_n$ are independent with

$$q_j|\mu,\sigma,\boldsymbol{y} \sim \Gamma\left(\frac{\nu+1}{2},\frac{1}{2}\left(\frac{(y_j-\mu)^2}{\sigma}+\nu\right)\right)$$

for $j \in \mathbb{N}_n$. Furthermore, making draws from $f(\sigma,\mu|\boldsymbol{q},\boldsymbol{y})$ can be done in two easy steps by noticing that

$$\sigma|\boldsymbol{q},\boldsymbol{y} \sim \mathrm{IG}\left(\frac{n+1}{2},\frac{q.\,\hat{\sigma}(\boldsymbol{q})}{2}\right) \quad \text{and} \quad \mu|\sigma,\boldsymbol{q},\boldsymbol{y} \sim \mathrm{N}\left(\hat{\mu}(\boldsymbol{q}),\frac{\sigma}{q.}\right),$$

where

$$q. = \sum_{j=1}^{n} q_j\,, \qquad \hat{\mu}(\boldsymbol{q}) = \frac{1}{q.}\sum_{j=1}^{n} y_j q_j\,, \qquad \hat{\sigma}(\boldsymbol{q}) = \frac{1}{q.}\sum_{j=1}^{n} q_j[y_j-\hat{\mu}(\boldsymbol{q})]^2\,,$$

and by $X \sim \mathrm{IG}(a,b)$ we mean that the density of X is

$$f_{\mathrm{IG}}(x;a,b) = \frac{b^a e^{-\frac{b}{x}}}{\Gamma(a)x^{a+1}}I_{(0,\infty)}(x); \tag{3.8}$$

i.e., X has the distribution of $1/Y$ when $Y \sim \Gamma(a,b)$. Thus, it is straightforward to simulate the chain defined by (3.5) in this case.

3.3 Meng and van Dyk's Extensions

The key idea underlying the DA algorithm is that it is easier to simulate the Markov chain Φ_{DA} than to sample directly from the target posterior. If, in addition, Φ_{DA} converges geometrically, then the algorithm is even more useful because of supplementary results, like central limit theorems and exact bounds for the length of the burn-in period. Unfortunately, there is a general conflict between

simplicity and speed. For example, the speed of the EM algorithm, which can be viewed as the deterministic analog of the DA algorithm, is inversely proportional to the amount of missing information. This suggests that a more clever augmentation may lead to a better DA algorithm. Indeed, this is the idea behind conditional and marginal augmentation, which are now described.

3.3.1 Conditional Augmentation

Suppose we are able to introduce a working parameter, say, $\alpha \in \mathbb{R}_+$, into the standard data augmentation density $f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\theta})$ yielding a family of densities, $\{g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha)\}_{\alpha \in \mathbb{R}_+}$, each of which is a candidate for use in the DA algorithm.[2] One way to do this is via a family of one-to-one transformations $D_\alpha : \mathcal{Q} \to \mathcal{Q}$, indexed by the working parameter α. That is, for each $\alpha \in \mathbb{R}_+$, there is an analog to the hierarchical model (3.6):

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta}, \alpha &\sim g(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta}, \alpha); \\ \boldsymbol{r}|\boldsymbol{\theta}, \alpha &\sim g(\boldsymbol{r}|\boldsymbol{\theta}, \alpha), \end{aligned} \tag{3.9}$$

where

$$g(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta}, \alpha) = f(\boldsymbol{y}|D_\alpha^{-1}(\boldsymbol{r}), \boldsymbol{\theta}), \tag{3.10}$$

and $g(\boldsymbol{r}|\boldsymbol{\theta}, \alpha)$ is obtained from $f(\boldsymbol{q}|\boldsymbol{\theta})$ by change of variable $\boldsymbol{r} = D_\alpha(\boldsymbol{q})$; i.e.,

$$g(\boldsymbol{r}|\boldsymbol{\theta}, \alpha) = f(D_\alpha^{-1}(\boldsymbol{r})|\boldsymbol{\theta}) \left| \left| \frac{\partial D_\alpha^{-1}(\boldsymbol{r})}{\partial \boldsymbol{r}} \right| \right|. \tag{3.11}$$

Usually, the original augmentation scheme, $f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\theta})$, is a member of the family with, say, $\alpha = 1$.

---

[2] To clarify the notation, we use f to denote densities associated with the standard DA algorithm and g for densities involved in van Dyk and Meng's extensions. Similarly, $\boldsymbol{q}$ refers to missing data in the standard DA and $\boldsymbol{r}$ denotes missing data in CA and MA.

Generally speaking, in order for each member of such a family to be a viable augmentation scheme, it must be the case that α is "invisible" in the marginal distribution of the observed data; i.e., we must have

$$\int_{\mathcal{Q}} g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha)\, d\boldsymbol{r} = f(\boldsymbol{y}|\boldsymbol{\theta}) \quad \forall\, \alpha \in \mathbb{R}_+\,,\ \forall\, \boldsymbol{y} \in \mathcal{Y},\ \forall\, \boldsymbol{\theta} \in \Theta. \tag{3.12}$$

Notice that under the hierarchy (3.9), the above condition is automatically satisfied, since for each $\alpha \in \mathbb{R}_+$

$$\begin{aligned} &\int_{\mathcal{Q}} g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha) d\boldsymbol{r} = \int_{\mathcal{Q}} g(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta}, \alpha) g(\boldsymbol{r}|\boldsymbol{\theta}, \alpha) d\boldsymbol{r} = \\ &\int_{\mathcal{Q}} f(\boldsymbol{y}|D_\alpha^{-1}(\boldsymbol{r}), \boldsymbol{\theta}) f(D_\alpha^{-1}(\boldsymbol{r})|\boldsymbol{\theta}) \left|\frac{\partial D_\alpha^{-1}(\boldsymbol{r})}{\partial \boldsymbol{r}}\right| d\boldsymbol{r} = \\ &\int_{\mathcal{Q}} f(\boldsymbol{y}|\boldsymbol{q}, \boldsymbol{\theta}) f(\boldsymbol{q}|\boldsymbol{\theta}) d\boldsymbol{q} = \int_{\mathcal{Q}} f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\theta}) d\boldsymbol{q} = f(\boldsymbol{y}|\boldsymbol{\theta}). \end{aligned} \tag{3.13}$$

Once the validity of the CA is verified through equation (3.12), we can fix α at any value that leads to relatively quickly converging Markov chain. For a specific value of α, the transition kernel of the corresponding Markov chain is

$$K_\alpha(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int_{\mathcal{Q}} g(\boldsymbol{\theta}|\boldsymbol{r}, \boldsymbol{y}, \alpha) g(\boldsymbol{r}|\boldsymbol{\theta}', \boldsymbol{y}, \alpha) d\boldsymbol{r}$$

and it is easy to show that $f(\boldsymbol{\theta}|\boldsymbol{y})$ remains its invariant density (since $g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha)$ is a legitimate data augmentation scheme). van Dyk and Meng (2001) provide guidance as to choosing the value of α that corresponds to the fastest converging Markov chain by establishing connections with the corresponding EM algorithm. We do not pursue these ideas any further since we are mainly interested in MA, which is presented next.

3.3.2 Marginal Augmentation

In MA, the working parameter α is incorporated into the augmented model through a proper working prior $\omega(\alpha)$. There are several different ways to use $g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha)$, $\pi(\boldsymbol{\theta})$ and $\omega(\alpha)$ to construct an MCMC algorithm with invariant

density $f(\boldsymbol{\theta}|\boldsymbol{y})$ (Meng and van Dyk, 1999). Here we describe the two methods that Meng and van Dyk (1999) call *Scheme 1* and *Scheme 2* (see also Liu et al., 1994).

Assume that α and $\boldsymbol{\theta}$ are a priori independent; i.e., $\omega(\alpha|\boldsymbol{\theta}) \equiv \omega(\alpha)\ \forall\ \boldsymbol{\theta} \in \Theta$. Then

$$g_\omega(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}) := \int_{\mathbb{R}_+} g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha)\, \omega(\alpha)\, d\alpha$$

is a viable augmentation scheme.[3] To see this notice that (3.12) and Fubini's theorem yield

$$\begin{aligned} &\int_{\mathcal{Q}} g_\omega(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}) d\boldsymbol{r} = \int_{\mathcal{Q}} \int_{\mathbb{R}_+} g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha) \omega(\alpha) d\alpha d\boldsymbol{r} \\ &= \int_{\mathbb{R}_+} \omega(\alpha) \int_{\mathcal{Q}} g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha) d\boldsymbol{r} d\alpha = f(\boldsymbol{y}|\boldsymbol{\theta}) \int_{\mathbb{R}_+} \omega(\alpha) d\alpha = f(\boldsymbol{y}|\boldsymbol{\theta}). \end{aligned} \tag{3.14}$$

Therefore, for each fixed proper prior, $\omega(\cdot)$, we can make approximate draws from the target posterior by simulating the Markov chain, $\Phi_{\text{MA}} = \{\boldsymbol{\theta}_i\}_{i=0}^{\infty}$, with Markov transition density

$$K_\omega(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int_{\mathcal{Q}} g_\omega(\boldsymbol{\theta}|\boldsymbol{r}, \boldsymbol{y})\, g_\omega(\boldsymbol{r}|\boldsymbol{\theta}', \boldsymbol{y})\, d\boldsymbol{r}\ ,$$

where

$$g_\omega(\boldsymbol{\theta}|\boldsymbol{r}, \boldsymbol{y}) := \frac{\int_{\mathbb{R}_+} g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha) \pi(\boldsymbol{\theta}) \omega(\alpha) d\alpha}{\int_\Theta \int_{\mathbb{R}_+} g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha) \pi(\boldsymbol{\theta}) \omega(\alpha) d\alpha d\boldsymbol{\theta}}$$

and

$$g_\omega(\boldsymbol{r}|\boldsymbol{\theta}, \boldsymbol{y}) := \frac{\int_{\mathbb{R}_+} g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha) \pi(\boldsymbol{\theta}) \omega(\alpha) d\alpha}{\int_{\mathcal{Q}} \int_{\mathbb{R}_+} g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha) \pi(\boldsymbol{\theta}) \omega(\alpha) d\alpha d\boldsymbol{r}}.$$

This is our version of Scheme 1. In Section 3.5 we show how it is related to Meng and van Dyk's work and we give more details on how the chain is actually run.

As expressed in the notation, the Markov transition density K_ω clearly depends on the choice of working prior $\omega(\alpha)$. Each different proper prior on α will result in a different version of Φ_{MA} with its own convergence properties. van Dyk

[3] It is important to recognize that $g_\omega(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta})$ and $f(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{\theta})$ are typically not the same.

and Meng (2001, Section 4) give a heuristic argument suggesting that Φ_{MA} should converge faster as the working prior, $\omega(\alpha)$, becomes more "diffuse." For example, consider the following parametric family of IG priors for α

$$\omega(\alpha;\delta) = \frac{\delta^{\delta}}{\Gamma(\delta)} \frac{e^{-\delta/\alpha}}{\alpha^{\delta+1}} I_{(0,\infty)}(\alpha) \,,$$

where $\delta > 0$. The "diffuseness" of this density increases as δ decreases[4]. Thus, it is tempting to consider what would happen to the convergence rate of Φ_{MA} under the improper prior $\omega(\alpha) = \alpha^{-1} I_{(0,\infty)}(\alpha)$. However, it is clear from Equation (3.14) that $g_\omega(\boldsymbol{r}|\boldsymbol{\theta}', \boldsymbol{y})$ will be improper when $\omega(\alpha)$ is improper and hence K_ω is not defined in that case. Therefore, we need a new sampling scheme when $\omega(\alpha)$ is improper. It is called Scheme 2 and is described in the next subsection.

3.3.3 Marginal Augmentation with Improper Prior

We now consider a different Markov chain in which α is actually one of the variables. Suppose that $\omega(\alpha)$ is proper and consider the joint density defined by

$$g(\boldsymbol{\theta}, \boldsymbol{r}, \alpha|\boldsymbol{y}) \propto g(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\theta}, \alpha)\, \omega(\alpha)\, \pi(\boldsymbol{\theta}) \,. \tag{3.15}$$

The main idea of Scheme 2 is to use a two variable Gibbs sampler based on the above density with components $(\boldsymbol{\theta}, \alpha)$ and $\boldsymbol{r}$. To see that it produces an approximate draw from $f(\boldsymbol{\theta}|\boldsymbol{y})$ notice that equation (3.12) shows that integrating $\boldsymbol{r}$ out of (3.15) yields

$$\int_{\mathcal{Q}} g(\boldsymbol{\theta}, \boldsymbol{r}, \alpha|\boldsymbol{y}) d\boldsymbol{r} = f(\boldsymbol{\theta}|\boldsymbol{y})\, m(\boldsymbol{y}) \omega(\alpha).$$

The reason for introducing this chain is not as an alternative to Φ_{MA}, but because, unlike Φ_{MA}, it may still be well-defined when the working prior is improper. When

[4] Notice that $\text{E}\alpha^2 < \infty$ iff $\delta > 2$

$\omega(\alpha)$ is improper we have

$$\int_{\mathbb{R}_+}\int_{\mathcal{Q}} g(\boldsymbol{\theta},\boldsymbol{r},\alpha|\boldsymbol{y})\,d\boldsymbol{r}\,d\alpha \propto f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\int_{\mathbb{R}_+}\omega(\alpha)\,d\alpha = \infty\,,$$

so the density in (3.15) is improper. Consequently, under an improper prior on α, the Gibbs sampler based on (3.15) with components $(\boldsymbol{\theta},\alpha)$ and $\boldsymbol{r}$ will not be positive recurrent (Hobert, 2001b). However, in certain situations, it is still possible to construct a useful MCMC algorithm using (3.15) when $\omega(\alpha)$ is improper. Indeed, we have already seen that for each fixed $(\boldsymbol{\theta},\alpha)\in\Theta\times\mathbb{R}_+$,

$$\int_{\mathcal{Q}} g(\boldsymbol{\theta},\boldsymbol{r},\alpha|\boldsymbol{y})\,d\boldsymbol{r} = f(\boldsymbol{\theta}|\boldsymbol{y})m(\boldsymbol{y})\omega(\alpha) < \infty\,. \tag{3.16}$$

Suppose further that for each fixed $\boldsymbol{r}\in\mathcal{Q}$,

$$\int_{\mathbb{R}_+}\int_{\Theta} g(\boldsymbol{\theta},\boldsymbol{r},\alpha|\boldsymbol{y})\,d\boldsymbol{\theta}\,d\alpha < \infty\,. \tag{3.17}$$

Then we can run a two-variable Gibbs sampler by iteratively sampling from the densities

$$\begin{aligned} g^*(\boldsymbol{r}|\boldsymbol{\theta},\alpha,\boldsymbol{y}) &= \frac{g(\boldsymbol{\theta},\boldsymbol{r},\alpha|\boldsymbol{y})}{\int_{\mathcal{Q}} g(\boldsymbol{\theta},\boldsymbol{r},\alpha|\boldsymbol{y})d\boldsymbol{r}} \quad \text{and} \\ g^*(\boldsymbol{\theta},\alpha|\boldsymbol{r},\boldsymbol{y}) &= \frac{g(\boldsymbol{\theta},\boldsymbol{r},\alpha|\boldsymbol{y})}{\int_{\mathbb{R}_+}\int_{\Theta} g(\boldsymbol{\theta},\boldsymbol{r},\alpha|\boldsymbol{y})d\boldsymbol{\theta}d\alpha} \end{aligned} \tag{3.18}$$

The "$*$"s are used to remind the reader that the "joint density" to which these conditionals correspond is improper. Following the notation of Meng and van Dyk (1999), we will call this method Scheme 2. Write the resulting Markov chain as $\Phi = \left\{\left(\boldsymbol{r}_i,(\boldsymbol{\theta}_i,\alpha_i)\right)\right\}_{i=0}^{\infty}$. It is easy to show that (3.15) is an (improper) invariant density for this chain, and hence (as mentioned above) Φ is *not positive recurrent*. Furthermore, while $\Phi_1 = \{\boldsymbol{r}_i\}_{i=0}^{\infty}$ and $\Phi_2 = \{(\boldsymbol{\theta}_i,\alpha_i)\}_{i=0}^{\infty}$ are themselves (non-positive recurrent) Markov chains (see; e.g. Liu et al., 1994), the marginal sequence $\Phi_{\text{MA}}^* = \{\boldsymbol{\theta}_i\}_{i=0}^{\infty}$ induced by this algorithm is not necessarily a Markov chain nor

is this sequence necessarily convergent. Before discussing Φ^*_{MA} in more detail we return again to Example 3.1.

Example 3.1 continued: One common transformation is $\boldsymbol{r} = D_\alpha(\boldsymbol{q}) = \alpha\boldsymbol{q}$, which corresponds to the alternative representation $t_1(\nu, \mu, \sigma) = \mu + \sqrt{\alpha}\sqrt{\sigma}z/\sqrt{r}$, where $z \sim \mathrm{N}(0,1)$ and $r \sim \alpha\chi^2_\nu/\nu$ with z and r independent. The corresponding MA algorithm is based on the following hierarchal model. Conditional on (μ, σ, α), let $(y_j, r_j), j = 1, \ldots, n$ be iid pairs satisfying

$$\begin{aligned} y_j|r_j, \mu, \sigma, \alpha &\sim \mathrm{N}\left(\mu, \frac{\alpha\sigma}{r_j}\right) \\ r_j|\mu, \sigma, \alpha &\sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2\alpha}\right), \end{aligned} \tag{3.19}$$

with $(\mu, \sigma, \alpha) \sim \pi(\mu, \sigma)\,\omega(\alpha)$, where $\omega(\alpha)$ is the improper prior density $\alpha^{-1}I_{(0,\infty)}(\alpha)$ and $\pi(\mu, \sigma) \propto \frac{1}{\sigma}$. The posterior "density" of $(\boldsymbol{r}, \mu, \sigma, \alpha)$ given $\boldsymbol{y}$ under the above hierarchy is

$$g(\mu, \sigma, \boldsymbol{r}, \alpha|\boldsymbol{y}) = \left[\prod_{j=1}^{n} g(y_j|r_j, \mu, \sigma, \alpha)\, g(r_j|\mu, \sigma, \alpha)\right] \pi(\mu, \sigma)\, \omega(\alpha)\,. \tag{3.20}$$

It is easy to check that (3.17) is satisfied, which makes it possible to define two legitimate conditional densities as follows

$$g^*(\boldsymbol{r}|\mu, \sigma, \alpha, \boldsymbol{y}) = \frac{g(\boldsymbol{r}, \mu, \sigma, \alpha|\boldsymbol{y})}{\int_{\mathbb{R}^n_+} g(\boldsymbol{r}, \mu, \sigma, \alpha|\boldsymbol{y})\, d\boldsymbol{r}}$$

and

$$g^*(\mu, \sigma, \alpha|\boldsymbol{r}, \boldsymbol{y}) = \frac{g(\boldsymbol{r}, \mu, \sigma, \alpha|\boldsymbol{y})}{\int_{\mathbb{R}_+}\int_{\mathbb{R}_+}\int_{\mathbb{R}} g(\boldsymbol{r}, \mu, \sigma, \alpha|\boldsymbol{y})\, d\mu\, d\sigma\, d\alpha}\,.$$

Sampling from these two densities is easy (Meng and van Dyk, 1999; Hobert, 2001a). For example, $g^*(\boldsymbol{r}|\mu, \sigma, \alpha, \boldsymbol{y})$ is just a product of gamma densities and it is

easy to verify that

$$\begin{aligned} g^*(\boldsymbol{\theta}, \alpha | \boldsymbol{r}, \boldsymbol{y}) &= g(\mu, \sigma, \alpha | \boldsymbol{r}, \boldsymbol{y}) = g(\alpha | \boldsymbol{r}, \boldsymbol{y}) g(\sigma | \alpha, \boldsymbol{r}, \boldsymbol{y}) g(\mu | \sigma, \alpha, \boldsymbol{r}, \boldsymbol{y}) \\ &= f_{\mathrm{IG}}\left(\alpha; \frac{n\nu}{2}, \frac{\nu r.}{2}\right) f_{\mathrm{IG}}\left(\sigma; \frac{n-1}{2}, \frac{\hat{\sigma}(\boldsymbol{r}) r.}{2\alpha}\right) f_{\mathrm{N}}\left(\mu; \hat{\mu}(\boldsymbol{r}), \frac{\alpha\sigma}{r.}\right) \\ &\propto \alpha^{-\frac{n(\nu+1)}{2}-1} e^{-\frac{r.}{2\alpha}\left\{\nu + \frac{\hat{\sigma}(\boldsymbol{r})}{\sigma} + \frac{[\mu - \hat{\mu}(\boldsymbol{r})]^2}{\sigma}\right\}} r.^{\frac{n(\nu+1)}{2}} [\hat{\sigma}(\boldsymbol{r})]^{\frac{n-1}{2}} \sigma^{-\frac{n+2}{2}}, \end{aligned}$$

where $f_{\mathrm{IG}}(\cdot;\cdot,\cdot)$ is defined in equation (3.8), and

$$f_{\mathrm{N}}(x; a, b) := \frac{1}{\sqrt{2\pi b}} e^{-\frac{(x-a)^2}{2b}}.$$

Therefore, we can apply the Gibbs sampler (by iterating between these two conditionals) to form the Markov chain $\Phi = \{(\boldsymbol{r}_i, (\mu_i, \sigma_i, \alpha_i))\}_{i=0}^{\infty}$, where $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{in})$. Also define

$$\Phi_1 = \{\boldsymbol{r}_i\}_{i=0}^{\infty} \quad \text{and} \quad \Phi_2 = \{(\mu_i, \sigma_i, \alpha_i)\}_{i=0}^{\infty},$$

which are themselves Markov chains.

The improper density (3.20) is invariant for Φ and consequently Φ, Φ_1, and Φ_2 are all non-positive recurrent with $f(\mu, \sigma | \boldsymbol{y})\, \omega(\alpha)$ being an invariant density for Φ_2. The fact that, under our choices for $\pi(\mu, \sigma)$ and $\omega(\alpha)$, the sequence

$$\Phi^*_{\mathrm{MA}} := \{(\mu_i, \sigma_i)\}_{i=0}^{\infty}$$

is a positive recurrent Markov chain with invariant density $f(\mu, \sigma | \boldsymbol{y})$ should follow by Lemma 1 from van Dyk and Meng (2001) or Theorem 4 from Liu and Wu (1999). That is, one can simulate the unstable Markov chain, Φ, and use the (μ, σ) component to explore the proper posterior density $f(\mu, \sigma | \boldsymbol{y})$. However, we have serious doubts about the validity of the results of van Dyk and Meng, and Liu and Wu. Consequently, we have developed our own proofs (see Sections 3.4 – 3.6).

We now discuss the conditions under which the sequence Φ^*_{MA} is a positive recurrent Markov chain. Recall that $\Phi_2 = \{(\boldsymbol{\theta}_i, \alpha_i)\}_{i=0}^{\infty}$ is a Markov chain with

Markov transition density given by

$$\int_{\mathcal{Q}} g^*(\boldsymbol{\theta}, \alpha|\boldsymbol{r}, \boldsymbol{y}) g^*(\boldsymbol{r}|\boldsymbol{\theta}', \alpha', \boldsymbol{y}) d\boldsymbol{r}.$$

A sufficient condition for $\Phi^*_{\text{MA}} = \{\boldsymbol{\theta}_i\}_{i=0}^{\infty}$ to be a Markov chain is that the integral

$$\int_{\mathcal{Q}} \int_{\mathbb{R}_+} g^*(\boldsymbol{\theta}, \alpha|\boldsymbol{r}, \boldsymbol{y}) g^*(\boldsymbol{r}|\boldsymbol{\theta}', \alpha', \boldsymbol{y}) d\alpha d\boldsymbol{r}. \tag{3.21}$$

does not depend on α'. Another sufficient condition under which the sequence Φ^*_{MA} is a Markov chain is given in Meng and van Dyk (1999). Here we establish a connection between their condition and the integral (3.21).

Lemma 3.1. *Assume the hierarchical structure of Equations* (3.9) - (3.11). *If* $\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta}, \alpha|D_{\alpha'}(\boldsymbol{q}), \boldsymbol{y}) d\alpha$ *does not depend on* α', *then* Φ^*_{MA} *is a Markov chain with Markov transition density*

$$K_{MA}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int_{\mathcal{Q}} \int_{\mathbb{R}_+} g^*(\boldsymbol{\theta}, \alpha|\boldsymbol{r}, \boldsymbol{y}) g^*(\boldsymbol{r}|\boldsymbol{\theta}', \alpha', \boldsymbol{y}) d\alpha d\boldsymbol{r}.$$

Proof. Notice that by Equations (3.18), (3.15), (3.10), and (3.11)

$$\begin{aligned} g^*(\boldsymbol{r}|\boldsymbol{\theta}', \alpha', \boldsymbol{y}) &:= \frac{g(\boldsymbol{\theta}', \boldsymbol{r}, \alpha'|\boldsymbol{y})}{\int_{\mathcal{Q}} g(\boldsymbol{\theta}', \tilde{\boldsymbol{r}}, \alpha'|\boldsymbol{y}) d\tilde{\boldsymbol{r}}} \\ &= \frac{g(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta}', \alpha') g(\boldsymbol{r}|\boldsymbol{\theta}', \alpha') \pi(\boldsymbol{\theta}') \omega(\alpha')}{\int_{\mathcal{Q}} g(\boldsymbol{y}|\boldsymbol{\theta}', \tilde{\boldsymbol{r}}, \alpha') g(\tilde{\boldsymbol{r}}|\alpha', \boldsymbol{\theta}') \pi(\boldsymbol{\theta}') \omega(\alpha') d\tilde{\boldsymbol{r}}} \\ &= \frac{f(\boldsymbol{y}|D^{-1}_{\alpha'}(\boldsymbol{r}), \boldsymbol{\theta}') f(D^{-1}_{\alpha'}(\boldsymbol{r})|\boldsymbol{\theta}') \left| \left| \frac{\partial D^{-1}_{\alpha'}(\boldsymbol{r})}{\partial \boldsymbol{r}} \right| \right|}{f(\boldsymbol{y}|\boldsymbol{\theta}')}. \end{aligned}$$

Now, after change of variable $\boldsymbol{q} = D^{-1}_{\alpha'}(\boldsymbol{r})$, we see that (3.21) can be written as

$$\begin{aligned} &\int_{\mathcal{Q}} \int_{\mathbb{R}_+} g^*(\boldsymbol{\theta}, \alpha|D_{\alpha'}(\boldsymbol{q}), \boldsymbol{y}) d\alpha \frac{f(\boldsymbol{y}|\boldsymbol{q}, \boldsymbol{\theta}') f(\boldsymbol{q}|\boldsymbol{\theta}')}{f(\boldsymbol{y}|\boldsymbol{\theta}')} d\boldsymbol{q} \\ &= \int_{\mathcal{Q}} \int_{\mathbb{R}_+} g^*(\boldsymbol{\theta}, \alpha|D_{\alpha'}(\boldsymbol{q}), \boldsymbol{y}) d\alpha f(\boldsymbol{q}|\boldsymbol{\theta}', \boldsymbol{y}) d\boldsymbol{q}, \end{aligned}$$

which combined with the assumption completes the proof. □

We were able to show that the condition of Lemma 3.1 is satisfied in a very general set-up (see Theorem 3.1). Here we conclude the section with an example

that demonstrates that the condition of Lemma 3.1 is satisfied at least for the univariate t model.

Example 3.1 continued: We now show that the condition of Lemma 3.1 is satisfied for the univariate t model. Using the definition of the IG density, we obtain that

$$g^*(\mu,\sigma|\boldsymbol{r},\boldsymbol{y}) \propto \left(r.\left\{\nu+\frac{\hat{\sigma}(\boldsymbol{r})}{\sigma}+\frac{[\mu-\hat{\mu}(\boldsymbol{r})]^2}{\sigma}\right\}\right)^{-\frac{n(\nu+1)}{2}} r.^{\frac{n(\nu+1)}{2}}[\hat{\sigma}(\boldsymbol{r})]^{\frac{n-1}{2}}\sigma^{-\frac{n+2}{2}}$$
$$= \left\{\nu+\frac{\hat{\sigma}(\boldsymbol{r})}{\sigma}+\frac{[\mu-\hat{\mu}(\boldsymbol{r})]^2}{\sigma}\right\}^{-\frac{n(\nu+1)}{2}}[\hat{\sigma}(\boldsymbol{r})]^{\frac{n-1}{2}}\sigma^{-\frac{n+2}{2}}.$$

Finally, since $\hat{\sigma}(\alpha'\boldsymbol{q}) = \hat{\sigma}(\boldsymbol{q})$ and $\hat{\mu}(\alpha'\boldsymbol{q}) = \hat{\mu}(\boldsymbol{q})$, it follows that the condition of Lemma 3.1 is satisfied.

For further details about MA with improper working priors, along with empirical evidence that it produces fast mixing MCMC algorithms, the reader is referred to van Dyk and Meng (2001). For more information on the use of non-positive recurrent Markov chains in MCMC algorithms, see Liu and Wu (1999), Hobert (2001b) and Lavine (2003).

Liu and Wu (1999) noticed that some of the properties of the Markov chain Φ^*_{MA} can be derived in a more general set-up if the family of transformations $\{D_\alpha(\boldsymbol{q})\}_{\alpha\in\mathbb{R}_+}$ has a specific group structure. Following their ideas, we will show, in the next section that, in general, the target density $f(\boldsymbol{\theta}|\boldsymbol{y})$ is an invariant density for Φ^*_{MA}, which provides peace of mind about the results for the multivariate t model.

3.4 Connection with Group Theory

We start with some fundamental definitions that can be found, for example, in Green (1987).

Definition 3.1. *If $\mathcal{A}$ is a nonempty set, then every function from $\mathcal{A}\times\mathcal{A}$ to $\mathcal{A}$ is called a binary operation in the set $\mathcal{A}$.*

Definition 3.2. *A set $\mathcal{A}$ together with a binary operation "$\cdot$" is called a group if the following properties are satisfied.*

(a) Associativity: $(\alpha_1 \cdot \alpha_2) \cdot \alpha_3 = \alpha_1 \cdot (\alpha_2 \cdot \alpha_3)$ *for all* $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{A}$.

(b) Identity: there exists an identity element $e \in \mathcal{A}$ *such that* $e \cdot \alpha = \alpha \cdot e = \alpha$ *for any* $\alpha \in \mathcal{A}$.

(c) Invertibility: for each $\alpha \in \mathcal{A}$, *there exists a unique inverse element* $\alpha^{-1} \in \mathcal{A}$ *such that* $\alpha \cdot \alpha^{-1} = \alpha^{-1} \cdot \alpha = e$.

The functions that preserve the group operation play a special role in group theory. Here is the formal definition of such functions.

Definition 3.3. *A function* $D : \mathcal{A} \to \mathcal{B}$ *between the groups* $(\mathcal{A}, \cdot)$ *and* $(\mathcal{B}, \cdot)$ *is called a homomorphism if*

(a) $D(\alpha_1 \cdot \alpha_2) = D(\alpha_1) \cdot D(\alpha_2)$ *for all* $\alpha_1, \alpha_2 \in \mathcal{A}$;

(b) $D(e_{\mathcal{A}}) = e_{\mathcal{B}}$.

We will now proceed to the main definition of this section, namely the notion of a *transformation group*, which can be viewed as a special case of the following general definition taken from Olver (1999).

Definition 3.4. *Let $\mathcal{Q}$ be a set. Define $\mathscr{D}(\mathcal{Q})$ to be the family of all one-to-one functions* $d : \mathcal{Q} \to \mathcal{Q}$.

$\mathscr{D}(\mathcal{Q})$ is a group in which the group operation is defined by composition of functions. The identity transformation plays the role of the identity element, and the inverse of a transformation is the usual functional inverse. If the space $\mathcal{Q}$ is equipped with some additional structure (like a topology), then the class of possible functions included in $\mathscr{D}(\mathcal{Q})$ may be restricted correspondingly. For example, when $\mathcal{Q} = \mathbb{R}^n$, $\mathscr{D}(\mathcal{Q})$ may contain all continuous invertible functions $d : \mathbb{R}^n \to \mathbb{R}^n$ (recall that the composition of two continuous functions is also continuous).

Definition 3.5. *For a given group $\mathcal{A}$, a transformation group, acting on a space $\mathcal{Q}$ is defined as any homomorphism $D : \mathcal{A} \to \mathcal{D}(\mathcal{Q})$ mapping $\mathcal{A}$ to the group of invertible functions on $\mathcal{Q}$, and is denoted by $(\mathcal{A}, \mathcal{Q}, D)$.*

In other words, for each element $\alpha \in \mathcal{A}$ there is a corresponding invertible function $D_\alpha : \mathcal{Q} \to \mathcal{Q}$. Because this identification is a homomorphism the following properties are satisfied:

(i) $D_{\alpha_1} \circ D_{\alpha_2}(q) := D_{\alpha_1}(D_{\alpha_2}(q)) \equiv D_{\alpha_1 \cdot \alpha_2}(q)$ for all $\alpha_1, \alpha_2 \in \mathcal{A}$;

(ii) $D_e(q) \equiv q$, where e is the identity element of $\mathcal{A}$;

(iii) $D_{\alpha^{-1}}(q) \equiv D_\alpha^{-1}(q)$ for any $\alpha \in \mathcal{A}$,

where (iii) is a consequence of (i) and (ii).

Definition 3.6. *For a fixed $q \in \mathcal{Q}$, the subset*

$$O_q := \{D_\alpha(q) : \alpha \in \mathcal{A}\} \subset \mathcal{Q}$$

is called the orbit of q.

Some authors use the term *group action* instead of transformation group (see, e.g., Eaton, 1989). Notice that the action of the group $\mathcal{A}$ on the space $\mathcal{Q}$ induces an equivalence relationship among the elements of $\mathcal{Q}$, namely two points are equivalent iff they are in the same orbit. To see this, notice that

(i) $q \in O_q$, since $D_e(q) = q$;

(ii) If $q_1 \in O_{q_2}$ then $q_2 \in O_{q_1}$, since there exists an $\alpha \in \mathcal{A}$ such that $q_1 = D_\alpha(q_2)$, implying that $D_{\alpha^{-1}}(q_1) = D_{\alpha^{-1}}(D_\alpha(q_2)) = D_{\alpha^{-1} \cdot \alpha}(q_2) = D_e(q_2) = q_2$;

(iii) If $q_1 \in O_{q_2}$ and $q_2 \in O_{q_3}$ then $q_1 \in O_{q_3}$, since there exist $\alpha_1, \alpha_2 \in \mathcal{A}$ such that $q_1 = D_{\alpha_1}(q_2) = D_{\alpha_1}(D_{\alpha_2}(q_3)) = D_{\alpha_1 \cdot \alpha_2}(q_3)$.

Thus, the space $\mathcal{Q}$ can be partitioned into disjoint sets, corresponding to the equivalence classes (the orbits), induced by the above equivalence relationship. By the axiom of choice, we can choose a representative element for each of these equivalence classes. Denote the set of the representative elements by $\mathcal{Z}$. Now, since

the equivalence classes partition $\mathcal{Q}$, each element $q \in \mathcal{Q}$ must belong to a unique equivalence class. Thus, if $z \in \mathcal{Z}$ is the corresponding representative of that class, then $q \in O_z$ and there exists $\alpha \in \mathcal{A}$ such that $q = D_\alpha(z)$ (see Example 3.2 below). In fact, it is easy to see that for any $q \in \mathcal{Q}$, there is unique $z \in \mathcal{Z}$ and *at least one* $\alpha \in \mathcal{A}$ such that $q = D_\alpha(z)$. This means that the map $D : \mathcal{A} \times \mathcal{Z} \to \mathcal{Q}$ defined by $D(\alpha, z) := D_\alpha(z)$ is onto. Suppose further that the index set $\mathcal{Z}$ can be chosen in such a way that the map $D(\alpha, z)$ is one-to-one and differentiable. When this is possible, the index set $\mathcal{Z}$ is called a *smooth cross-section*. Usually, the smooth cross-section, $\mathcal{Z}$, has lower dimension than the space $\mathcal{Q}$ and it is convenient to write $\mathcal{Z}$ as the image of another space $\mathcal{R}$ via the one-to-one transformation Z; i.e. $\mathcal{Z} = Z(\mathcal{R})$. Here is an example.

Example 3.2: Let $\mathcal{A}$ be the group of all positive real numbers, $\mathbb{R}_+$, with product as the group binary operation. Obviously, $1 \in \mathbb{R}_+$ is the identity element and the reciprocal $\frac{1}{\alpha} = \alpha^{-1}$ is the unique inverse of $\alpha \in \mathbb{R}_+$. Also let $\mathcal{Q} = \mathbb{R}^n_+$ and for any $\alpha \in \mathbb{R}_+$ define $D_\alpha : \mathbb{R}^n_+ \to \mathbb{R}^n_+$ by

$$D_\alpha(\boldsymbol{q}) := \alpha \boldsymbol{q},$$

which is a one-to-one transformation satisfying the conditions of Definition 3.5. The structure $\left(\mathbb{R}_+, \mathbb{R}^n_+, \{D_\alpha\}_{\alpha \in \mathbb{R}_+}\right)$ is called the *scale transformation group*. Notice that the set of all orbits is the set of all lines in the first quadrant that go through the origin. They do not intersect since the point $\mathbf{0}$ is not in $\mathbb{R}^n_+$. A smooth cross-section is the part of the unit sphere that is restricted to the first quadrant; i.e.,

$$\mathcal{Z} := \{(q_1, \ldots, q_n) \in \mathbb{R}^n_+ : q_1^2 + \cdots + q_n^2 = 1\}.$$

Furthermore, any point $\boldsymbol{q} \in \mathbb{R}^n_+$ can be represented as $\boldsymbol{q} = D_\alpha(\boldsymbol{z}) = \alpha \boldsymbol{z}$, where $\alpha \in \mathbb{R}_+$ and $\boldsymbol{z} \in \mathcal{Z}$ with $\alpha := \|\boldsymbol{q}\|$ and $\boldsymbol{z} := \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|}$. Finally, any point

$\boldsymbol{z} = (z_1, \ldots, z_n) \in \mathcal{Z}$ can be written as

$$
\begin{aligned}
z_1 &= \cos r_1 \cos r_2 \ldots \cos r_{n-3} \cos r_{n-2} \cos r_{n-1} \\
z_2 &= \cos r_1 \cos r_2 \ldots \cos r_{n-3} \cos r_{n-2} \sin r_{n-1} \\
z_3 &= \cos r_1 \cos r_2 \ldots \cos r_{n-3} \sin r_{n-2} \\
&\vdots \\
z_n &= \sin r_1,
\end{aligned}
$$

where $\boldsymbol{r} := (r_1, \ldots, r_{n-1})^\top \in \mathcal{R} := (0, \frac{\pi}{2})^{n-1}$.

We will now show that for the scale transformation group from Example 3.2 with improper prior $\frac{1}{\alpha}$, the MA algorithm works, by proving that Φ^*_{MA} is a Markov chain and that the target density $f(\boldsymbol{\theta}|\boldsymbol{y})$ is invariant for the transition kernel K_{MA}. Liu and Wu (1999) claim to prove this result for any transformation group as long as the improper prior, $\omega(\alpha)$, corresponds to the density of the Haar measure of the family of transformations $D_\alpha(q)$. Unfortunately, they did not provide adequate definitions and consequently it is very hard to follow the proof of their theorem, so we provide an alternative proof for the scale transformation group and invite the reader to attempt to follow the discourse in Liu and Wu (1999).

Theorem 3.1. *Assume the hierarchical structure of Equations* (3.9) - (3.11) *and consider MA with improper prior* $\omega(\alpha) = \frac{1}{\alpha}$. *If the transformation is given by* $D_\alpha(\boldsymbol{q}) = \alpha \boldsymbol{q}$, *then* Φ^*_{MA} *is a Markov chain and*

$$\int_\Theta K_{MA}(\boldsymbol{\theta}|\boldsymbol{\theta}') f(\boldsymbol{\theta}'|\boldsymbol{y}) d\boldsymbol{\theta}' = f(\boldsymbol{\theta}|\boldsymbol{y}).$$

Proof. We will establish that Φ^*_{MA} is a Markov chain on our way to showing that $f(\boldsymbol{\theta}|\boldsymbol{y})$ is an invariant density for $K_{\mathrm{MA}}(\boldsymbol{\theta}|\boldsymbol{\theta}')$. First notice that based on the proof of Lemma 3.1, we can write

$$K_{\mathrm{MA}}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int_{\mathcal{Q}} \int_{\mathbb{R}_+} g^*(\boldsymbol{\theta}, \alpha | D_{\alpha'}(\boldsymbol{q}), \boldsymbol{y}) d\alpha f(\boldsymbol{q}|\boldsymbol{\theta}', \boldsymbol{y}) d\boldsymbol{q}.$$

Then

$$\begin{aligned}
&\int_{\Theta} K_{\mathrm{MA}}(\boldsymbol{\theta}|\boldsymbol{\theta}')f(\boldsymbol{\theta}'|\boldsymbol{y})d\boldsymbol{\theta}' \\
&= \int_{\Theta}\int_{\mathcal{Q}}\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta},\alpha|D_{\alpha'}(\boldsymbol{q}),\boldsymbol{y})d\alpha f(\boldsymbol{q}|\boldsymbol{\theta}',\boldsymbol{y})f(\boldsymbol{\theta}'|\boldsymbol{y})d\boldsymbol{q}d\boldsymbol{\theta}' \\
&= \int_{\mathcal{Q}}\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta},\alpha|D_{\alpha'}(\boldsymbol{q}),\boldsymbol{y})d\alpha \int_{\Theta} f(\boldsymbol{q},\boldsymbol{\theta}'|\boldsymbol{y})d\boldsymbol{\theta}'d\boldsymbol{q} \\
&= \int_{\mathcal{Q}}\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta},\alpha|D_{\alpha'}(\boldsymbol{q}),\boldsymbol{y})d\alpha f(\boldsymbol{q}|\boldsymbol{y})d\boldsymbol{q},
\end{aligned}$$

where

$$f(\boldsymbol{q}|\boldsymbol{y}) := \int_{\Theta} f(\boldsymbol{q},\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}.$$

Therefore, we need to show that

$$\int_{\mathcal{Q}}\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta},\alpha|D_{\alpha'}(\boldsymbol{q}),\boldsymbol{y})d\alpha f(\boldsymbol{q}|\boldsymbol{y})d\boldsymbol{q} = f(\boldsymbol{\theta}|\boldsymbol{y}).$$

Recall that we can write $\boldsymbol{q} = D(s,\boldsymbol{z}) = D(s,Z(\boldsymbol{r}))$, where the mapping $D : \mathcal{A}\times\mathcal{Z}\to\mathcal{Q}$ is one-to-one and differentiable. Thus, we can make change of variables $\boldsymbol{q} = D_s(Z(\boldsymbol{r}))$ with Jacobian J equal to the absolute value of the determinant of the $n\times n$ matrix

$$\left(\frac{\partial D_s(Z(\boldsymbol{r}))}{\partial s}\ \frac{\partial D_s(Z(\boldsymbol{r}))}{\partial \boldsymbol{r}}\right).$$

In our case (the scale transformation group), $D_s(Z(\boldsymbol{r})) = sZ(\boldsymbol{r})$ and thus $\frac{\partial D_s(Z(\boldsymbol{r}))}{\partial s} = Z(\boldsymbol{r})$. Therefore, $J = s^{n-1}L(\boldsymbol{r})$, where $L : (0,\frac{\pi}{2})^{n-1}\to\mathbb{R}$ is a function of $\boldsymbol{r}$ only (this can be seen from the definition of determinant and the fact that s is a scalar in all of the elements of the matrix except those in the first column). Hence, we need to prove

$$\int_{\mathbb{R}_+}\int_{\mathcal{R}}\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta},\alpha|D_{\alpha'}(D_s(Z(\boldsymbol{r}))),\boldsymbol{y})d\alpha f(D_s(Z(\boldsymbol{r}))|\boldsymbol{y})s^{n-1}L(\boldsymbol{r})d\boldsymbol{r}ds = f(\boldsymbol{\theta}|\boldsymbol{y}),$$

which is equivalent to

$$\int_{\mathbb{R}_+}\int_{\mathcal{R}}\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta},\alpha|\alpha' sZ(\boldsymbol{r})),\boldsymbol{y})d\alpha f(sZ(\boldsymbol{r})|\boldsymbol{y})s^{n-1}L(\boldsymbol{r})d\boldsymbol{r}ds = f(\boldsymbol{\theta}|\boldsymbol{y}).$$

Now by definition (see equation (3.18)),

$$g^*(\boldsymbol{\theta},\alpha|\alpha' sZ(\boldsymbol{r}),\boldsymbol{y}) = \frac{g(\boldsymbol{y},\alpha' sZ(\boldsymbol{r})|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta})\omega(\alpha)}{\int_{\mathbb{R}_+}\int_\Theta g(\boldsymbol{y},\alpha' sZ(\boldsymbol{r})|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta})\omega(\alpha)d\boldsymbol{\theta}d\alpha}$$
$$= \frac{f(\boldsymbol{y},D_\alpha^{-1}(\alpha' sZ(\boldsymbol{r})),|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\omega(\alpha)J_\alpha(\alpha' sZ(\boldsymbol{r}))}{\int_{\mathbb{R}_+}\int_\Theta f(\boldsymbol{y},D_\alpha^{-1}(\alpha' sZ(\boldsymbol{r}))|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\omega(\alpha)J_\alpha(\alpha' sZ(\boldsymbol{r}))d\boldsymbol{\theta}d\alpha}$$

and since $J_\alpha(\boldsymbol{q}) := \left|\left|\frac{\partial D_\alpha^{-1}(\boldsymbol{r})}{\partial \boldsymbol{r}}\right|\right| = \frac{1}{\alpha^n}$, we have

$$g^*(\boldsymbol{\theta},\alpha|\alpha' sZ(\boldsymbol{r}),\boldsymbol{y}) = \frac{f(\boldsymbol{\theta},\alpha^{-1}\alpha' sZ(\boldsymbol{r})|\boldsymbol{y})m(\boldsymbol{y})\omega(\alpha)J_\alpha(\alpha' sZ(\boldsymbol{r}))}{\int_{\mathbb{R}_+} f(\alpha^{-1}\alpha' sZ(\boldsymbol{r})|\boldsymbol{y})m(\boldsymbol{y})\omega(\alpha)J_\alpha(\alpha' sZ(\boldsymbol{r}))d\alpha}$$
$$= \frac{f(\alpha^{-1}\alpha' sZ(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})\alpha^{-n-1}}{\int_{\mathbb{R}_+} f(\alpha^{-1}\alpha' sZ(\boldsymbol{r})|\boldsymbol{y})\alpha^{-n-1}d\alpha}.$$

Another change of variable $\beta = \alpha^{-1}\alpha' s$ with Jacobian $\frac{\alpha' s}{\beta^2}$ gives us

$$\int_{\mathbb{R}_+} f(\alpha^{-1}\alpha' sZ(\boldsymbol{r})|\boldsymbol{y})\alpha^{-n-1}d\alpha = \int_{\mathbb{R}_+} f(\beta Z(\boldsymbol{r})|\boldsymbol{y})\left(\frac{\beta}{\alpha' s}\right)^{n+1}\frac{\alpha' s}{\beta^2}d\beta$$
$$= \left(\frac{1}{\alpha' s}\right)^n \int_{\mathbb{R}_+} f(\beta Z(\boldsymbol{r})|\boldsymbol{y})\beta^{n-1}d\beta$$
$$= \left(\frac{1}{\alpha' s}\right)^n \frac{1}{L(\boldsymbol{r})}\int_{\mathbb{R}_+} f(\beta Z(\boldsymbol{r})|\boldsymbol{y})\beta^{n-1}L(\boldsymbol{r})d\beta$$
$$= \left(\frac{1}{\alpha' s}\right)^n \frac{h(Z(\boldsymbol{r})|\boldsymbol{y})}{L(\boldsymbol{r})},$$

where

$$h(Z(\boldsymbol{r})|\boldsymbol{y}) := \int_{\mathbb{R}_+} u(s,Z(\boldsymbol{r})|\boldsymbol{y})ds$$

and $u(s,Z(\boldsymbol{r})|\boldsymbol{y})$ is obtained from $f(\boldsymbol{q}|\boldsymbol{y})$ after change of variable $\boldsymbol{q} = D_s(Z(\boldsymbol{r})) = sZ(\boldsymbol{r})$. Therefore,

$$g^*(\boldsymbol{\theta},\alpha|\alpha' sZ(\boldsymbol{r}),\boldsymbol{y}) = \frac{f(\alpha^{-1}\alpha' sZ(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})\alpha^{-n-1}}{\left(\frac{1}{\alpha' s}\right)^n \frac{h(Z(\boldsymbol{r}),\boldsymbol{y})}{L(\boldsymbol{r})}}.$$

Similarly, we can make the change of variable to obtain that

$$\int_{\mathbb{R}_+} f(\alpha^{-1}\alpha' sZ(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})\alpha^{-n-1}d\alpha = \left(\frac{1}{\alpha' s}\right)^n \frac{h(Z(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})}{L(\boldsymbol{r})},$$

where

$$h(Z(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y}) := \int_{\mathbb{R}_+} u(s, Z(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})ds$$

and $u(s, Z(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})$ is obtained from $f(\boldsymbol{q},\boldsymbol{\theta}|\boldsymbol{y})$ after change of variable $\boldsymbol{q} = D_s(Z(\boldsymbol{r})) = sZ(\boldsymbol{r})$. Thus,

$$\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta},\alpha|D_{\alpha'}(D_s(Z(\boldsymbol{r}))),\boldsymbol{y})d\alpha = \frac{h(Z(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})}{h(Z(\boldsymbol{r})|\boldsymbol{y})}$$

which does not depend on α'. Thus, by Lemma 3.1, the sequence $\Phi^*_{\mathrm{M}\Lambda}$ is a Markov chain. Notice that all the results after the change of variable $\beta = \alpha^{-1}\alpha' s$ depend heavily on the choice of prior $\omega(\alpha)$. It is our belief that this is where the Haar measure plays a special role in the proof. However, it still remains to be investigated rigorously what happens in the general case.

Now, in order to show that $f(\boldsymbol{\theta}|\boldsymbol{y})$ is the invariant density for the Markov chain $\Phi^*_{\mathrm{M}\Lambda}$, we have to prove that

$$\int_{\mathcal{R}} \frac{h(Z(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})}{h(Z(\boldsymbol{r})|\boldsymbol{y})} \int_{\mathbb{R}_+} f(sZ(\boldsymbol{r})|\boldsymbol{y})s^{n-1}L(\boldsymbol{r})dsd\boldsymbol{r} = f(\boldsymbol{\theta}|\boldsymbol{y}),$$

or equivalently (by the definition of $h(Z(\boldsymbol{r})|\boldsymbol{y})$),

$$\int_{\mathcal{R}} h(Z(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{r} = f(\boldsymbol{\theta}|\boldsymbol{y}). \tag{3.22}$$

But

$$h(Z(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y}) := \int_{\mathbb{R}_+} u(s, Z(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})ds = \int_{\mathbb{R}_+} f(sZ(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})s^{n-1}L(\boldsymbol{r})ds.$$

Therefore, (3.22) is equivalent to

$$\int_{\mathcal{R}}\int_{\mathbb{R}_+} f(sZ(\boldsymbol{r}),\boldsymbol{\theta}|\boldsymbol{y})s^{n-1}L(\boldsymbol{r})dsd\boldsymbol{r} = f(\boldsymbol{\theta}|\boldsymbol{y}).$$

Finally, the inverse transformation $\boldsymbol{q} = sZ(\boldsymbol{r})$ shows that the above is equivalent to

$$\int_{\mathcal{Q}} f(\boldsymbol{\theta},\boldsymbol{q}|\boldsymbol{y})d\boldsymbol{q} = f(\boldsymbol{\theta}|\boldsymbol{y}),$$

which we already know is satisfied. □

When the conditions of Theorem 3.1 are satisfied, one can simulate the non-positive recurrent chain Φ and actually use the marginal chain Φ^*_{MA} to explore the proper, but intractable posterior $f(\boldsymbol{\theta}|\boldsymbol{y})$. Moreover, this phenomenon is more than just a theoretical curiosity. Indeed, there is substantial empirical evidence suggesting that Φ^*_{MA} converges faster than Φ_{MA} and much faster than Φ_{DA}. It is not yet fully understood why an algorithm that is based on a non-positive recurrent Markov chain converges faster than related algorithms based on positive recurrent chains.

3.5 Understanding the Artwork of van Dyk and Meng

In this section we will show that the descriptions van Dyk and Meng give to define their Markov chains agree with our definitions. We start with Scheme 1; i.e., the conditional densities used in the expressions that follow are based on the joint density $g(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{\theta},\alpha)g(\boldsymbol{r}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta})\omega(\alpha)$ that is now based on a proper $\omega(\alpha)$. First note that

$$g(\boldsymbol{r}|\boldsymbol{\theta}',\boldsymbol{y}) := \int_{\mathbb{R}_+} g(\boldsymbol{r},\alpha|\boldsymbol{\theta}',\boldsymbol{y})d\alpha = \int_{\mathbb{R}_+} g(\boldsymbol{r}|\boldsymbol{\theta}',\alpha,\boldsymbol{y})g(\alpha|\boldsymbol{\theta}',\boldsymbol{y})d\alpha.$$

But

$$\begin{aligned} g(\alpha|\boldsymbol{\theta}',\boldsymbol{y}) &= \int_{\mathcal{Q}} g(\boldsymbol{r},\alpha|\boldsymbol{\theta}',\boldsymbol{y})d\boldsymbol{r} \\ &= \int_{\mathcal{Q}} \frac{g(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{\theta}',\alpha)g(\boldsymbol{r}|\boldsymbol{\theta}',\alpha)\pi(\boldsymbol{\theta}')\omega(\alpha)}{\int_{\mathbb{R}_+}\int_{\mathcal{Q}} g(\boldsymbol{y}|\tilde{\boldsymbol{r}},\boldsymbol{\theta}',\alpha)g(\tilde{\boldsymbol{r}}|\boldsymbol{\theta}',\alpha)\pi(\boldsymbol{\theta}')\omega(\alpha)d\tilde{\boldsymbol{r}}d\alpha}d\boldsymbol{r} \\ &= \frac{\int_{\mathcal{Q}} g(\boldsymbol{y},\boldsymbol{r}|\boldsymbol{\theta}',\alpha)d\boldsymbol{r}\omega(\alpha)}{\int_{\mathbb{R}_+}\int_{\mathcal{Q}} g(\boldsymbol{y},\tilde{\boldsymbol{r}}|\boldsymbol{\theta}',\alpha)\omega(\alpha)d\tilde{\boldsymbol{r}}d\alpha} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}')\omega(\alpha)}{f(\boldsymbol{y}|\boldsymbol{\theta}')\int_{\mathbb{R}_+}\omega(\alpha)d\alpha} = \omega(\alpha). \end{aligned}$$

Therefore,

$$g(\boldsymbol{r}|\boldsymbol{\theta}',\boldsymbol{y}) = \int_{\mathbb{R}_+} g(\boldsymbol{r}|\boldsymbol{\theta}',\alpha,\boldsymbol{y})\,\omega(\alpha)\,d\alpha\,. \tag{3.23}$$

Hence, K_ω can be alternatively represented as

$$\begin{aligned} & K_\omega(\boldsymbol{\theta}|\boldsymbol{\theta}') \\ &= \int_{\mathcal{Q}} \left[\int_{\mathbb{R}_+} g(\boldsymbol{\theta}|\boldsymbol{r}, \alpha'', \boldsymbol{y}) g(\alpha''|\boldsymbol{r}, \boldsymbol{y}) d\alpha'' \right] \left[\int_{\mathbb{R}_+} g(\boldsymbol{r}|\boldsymbol{\theta}', \alpha', \boldsymbol{y}) \omega(\alpha') d\alpha' \right] d\boldsymbol{r}. \end{aligned} \tag{3.24}$$

In Meng and van Dyk (1999) page 309 and van Dyk and Meng (2001) Lemma 1, the authors claim that the Markov chain corresponding to Scheme 1 (according to us, it has transition kernel (3.24)) can be simulated by performing the following steps: first draw α' from the prior $\omega(\alpha')$, and then draw $\boldsymbol{r}$ from $g(\boldsymbol{r}|\boldsymbol{\theta}', \alpha', \boldsymbol{y})$; second draw α'' from the posterior $g(\alpha''|\boldsymbol{r}, \boldsymbol{y})$, and then draw $\boldsymbol{\theta}$ from $g(\boldsymbol{\theta}|\boldsymbol{r}, \alpha'', \boldsymbol{y})$. In other words, the following steps must be performed:

1. Draw $\boldsymbol{q}$ from $f(\boldsymbol{q}|\boldsymbol{\theta}', \boldsymbol{y})$;
2. Draw α' from $\omega(\alpha')$;
3. Take $\boldsymbol{r} = D_{\alpha'}(\boldsymbol{q})$;
4. Draw $\boldsymbol{\theta}$ from $g(\boldsymbol{\theta}|\boldsymbol{r}, \boldsymbol{y})$.

However, Meng and van Dyk did not provide the transition kernel for their Markov chain which makes it harder to understand their idea. To see that van Dyk and Meng's and our description agree, first notice the simple fact that if we need a draw from $f(y) = \int_{\mathcal{X}} f(x, y) dx$, we can first get $x \sim f(x)$ and then $y \sim f(y|x)$, and the resulting y is marginally from $f(y)$. Now notice that conditionally on α' , the variable $\boldsymbol{r}$ obtained in step 3 has density $f(D_{\alpha'}^{-1}(\boldsymbol{r})|\boldsymbol{\theta}', \boldsymbol{y}) J_{\alpha'}(\boldsymbol{r})$, where

$$J_{\alpha'}(\boldsymbol{r}) = \left| \left| \frac{\partial D_{\alpha'}^{-1}(\boldsymbol{r})}{\partial \boldsymbol{r}} \right| \right|$$

Thus, the result of steps 1, 2 and 3 is a draw from $\int_{\mathbb{R}_+} f(D_{\alpha'}^{-1}(\boldsymbol{r})|\boldsymbol{\theta}', \boldsymbol{y}) J_{\alpha'}(\boldsymbol{r}) \omega(\alpha') d\alpha'$. But

$$f(\boldsymbol{q}|\boldsymbol{\theta}', \boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{q}, \boldsymbol{\theta}') f(\boldsymbol{q}|\boldsymbol{\theta}')}{f(\boldsymbol{y}|\boldsymbol{\theta}')}$$

and therefore

$$f(D_{\alpha'}^{-1}(\boldsymbol{r})|\boldsymbol{\theta}',\boldsymbol{y})J_{\alpha'}(\boldsymbol{r}) = \frac{f(\boldsymbol{y}|D_{\alpha'}^{-1}(\boldsymbol{r}),\boldsymbol{\theta}')f(D_{\alpha'}^{-1}(\boldsymbol{r})|\boldsymbol{\theta}')J_{\alpha'}(\boldsymbol{r})}{f(\boldsymbol{y}|\boldsymbol{\theta}')}$$
$$= \frac{g(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{\theta}',\alpha')g(\boldsymbol{r}|\boldsymbol{\theta}',\alpha')}{f(\boldsymbol{y}|\boldsymbol{\theta}')} = g(\boldsymbol{r}|\boldsymbol{\theta}',\alpha',\boldsymbol{y}). \tag{3.25}$$

Thus, the final result after step 3 is a draw from

$$\int_{\mathbb{R}_+} f(D_{\alpha'}^{-1}(\boldsymbol{r})|\boldsymbol{\theta}',\boldsymbol{y})J_{\alpha'}(\boldsymbol{r})\omega(\alpha')d\alpha' = \int_{\mathbb{R}_+} g(\boldsymbol{r}|\boldsymbol{\theta}',\alpha',\boldsymbol{y})\omega(\alpha')d\alpha'.$$

This combined with step 4 makes it crystal clear that the two definitions of the transition kernel under Scheme 1 agree.

We now briefly discuss the same issue but for Scheme 2. As noted in van Dyk and Meng (2001), K_{MA} can be described similarly to what we did in Scheme 1. For any fixed $\alpha' \in \mathbb{R}_+$ perform the following steps:

1. Draw $\boldsymbol{q}$ from $f(\boldsymbol{q}|\boldsymbol{\theta}',\boldsymbol{y})$;
2. Take $\boldsymbol{r} = D_{\alpha'}(\boldsymbol{q})$;
3. Draw $\boldsymbol{\theta}$ from $\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta},\alpha|\boldsymbol{r},\boldsymbol{y})d\alpha$.

Similarly to (3.25), the result of steps 1 and 2 above is a draw from $g^*(\boldsymbol{r}|\boldsymbol{\theta}',\alpha',\boldsymbol{y})$. Therefore, our definition for Scheme 2 agrees with the above.

We now address the question about the stationary distribution of Φ_{MA}^*. There are problems with van Dyk and Meng's results concerning the invariance of $f(\boldsymbol{\theta}|\boldsymbol{y})$ for Φ_{MA}^* and these are described now. Since these authors never defined their Markov chains with transition kernels, we start by first restating their main result. As can be seen afterwards, this result might not even be true. The following theorem is an attempt to rigorously state Lemma 1 from van Dyk and Meng (2001).

Theorem 3.2. *Assume the condition of Lemma 3.1 is satisfied and let* $K_{MA}(\boldsymbol{\theta}|\boldsymbol{\theta}')$ *be the transition kernel of* Φ_{MA}^* *when implementing Scheme 2 under an improper prior* $\omega(\alpha)$ *and let* $k_m(\boldsymbol{\theta}|\boldsymbol{\theta}')$ *be the transition kernel of* Φ_{MA} *when*

implementing Scheme 1 under the proper prior $\omega_m(\alpha)$, where $m \in \mathbb{N}$. Define

$$g_m(\boldsymbol{\theta}, \alpha|\boldsymbol{r}, \boldsymbol{y}) := \frac{g(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta}, \alpha)g(\boldsymbol{r}|\boldsymbol{\theta}, \alpha)\pi(\boldsymbol{\theta})\omega_m(\alpha)}{\int_\Theta \int_{\mathbb{R}_+} g(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta}, \alpha)g(\boldsymbol{r}|\boldsymbol{\theta}, \alpha)\pi(\boldsymbol{\theta})\omega_m(\alpha)d\alpha d\boldsymbol{\theta}}$$

and, similarly,

$$g_m(\boldsymbol{r}, \alpha|\boldsymbol{\theta}, \boldsymbol{y}) := \frac{g(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta}, \alpha)g(\boldsymbol{r}|\boldsymbol{\theta}, \alpha)\pi(\boldsymbol{\theta})\omega_m(\alpha)}{\int_\mathcal{Q} \int_{\mathbb{R}_+} g(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\theta}, \alpha)g(\boldsymbol{r}|\boldsymbol{\theta}, \alpha)\pi(\boldsymbol{\theta})\omega_m(\alpha)d\alpha d\boldsymbol{r}}$$

If the limit $\lim_{m\to\infty} g_m(\boldsymbol{\theta}, \alpha|\boldsymbol{r}, \boldsymbol{y})$ exists and

$$\lim_{m\to\infty} g_m(\boldsymbol{\theta}, \alpha|\boldsymbol{r}, \boldsymbol{y}) = g^*(\boldsymbol{\theta}, \alpha|\boldsymbol{r}, \boldsymbol{y}),$$

where $g^*(\boldsymbol{\theta}, \alpha|\boldsymbol{r}, \boldsymbol{y})$ is calculated using the improper prior $\omega(\alpha)$, then

$$k_m(\boldsymbol{\theta}|\boldsymbol{\theta}') \to K_{MA}(\boldsymbol{\theta}|\boldsymbol{\theta}') \text{ as } m \to \infty,$$

implying that the stationary distribution of Φ^*_{MA} is $f(\boldsymbol{\theta}|\boldsymbol{y})$.

The fact that $k_m(\boldsymbol{\theta}|\boldsymbol{\theta}') \to K_{\mathrm{MA}}(\boldsymbol{\theta}|\boldsymbol{\theta}')$ as $m \to \infty$ implies that $f(\boldsymbol{\theta}|\boldsymbol{y})$ is invariant for K_{MA}, follows from Liu and Wu's Lemma 1, which we restate here for reference.

Lemma 3.2 (Liu and Wu). *Suppose that $K_m(y|x)$ is a sequence of Markov transition functions, all having $\pi(x)$ as invariant density. If $K(y|x) := \lim_{m\to\infty} K_m(y|x)$ is a proper transition function, then π is an invariant density of K.*

If true, Theorem 3.2 would provide justification for the use of the improper prior $\omega(\alpha)$. Unfortunately, the "proof" of van Dyk and Meng has some gaps. More specifically, notice the following

$$\begin{aligned} k_m(\boldsymbol{\theta}|\boldsymbol{\theta}') &:= \int_\mathcal{Q} g_m(\boldsymbol{\theta}|\boldsymbol{r}, \boldsymbol{y})g_m(\boldsymbol{r}|\boldsymbol{\theta}', \boldsymbol{y})d\boldsymbol{r} \\ &= \int_\mathcal{Q} \int_{\mathbb{R}_+} g_m(\boldsymbol{\theta}, \alpha|\boldsymbol{r}, \boldsymbol{y})g_m(\boldsymbol{r}|\boldsymbol{\theta}', \boldsymbol{y})d\alpha d\boldsymbol{r} \end{aligned}$$

and by equation (3.23)

$$
\begin{aligned}
k_m(\boldsymbol{\theta}|\boldsymbol{\theta}') &= \int_{\mathcal{Q}}\int_{\mathbb{R}_+} g_m(\boldsymbol{\theta},\alpha|\boldsymbol{r},\boldsymbol{y}) \int_{\mathbb{R}_+} g(\boldsymbol{r}|\boldsymbol{\theta}',\alpha',\boldsymbol{y})\omega_m(\alpha')d\alpha' d\alpha d\boldsymbol{r} \\
&= \int_{\mathcal{Q}}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \frac{g(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{\theta},\alpha)g(\boldsymbol{r}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta})\omega_m(\alpha)}{\int_\Theta\int_{\mathbb{R}_+} g(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{\theta},\tilde{\alpha})g(\boldsymbol{r}|\boldsymbol{\theta},\tilde{\alpha})\pi(\boldsymbol{\theta})\omega_m(\tilde{\alpha})d\tilde{\alpha}d\boldsymbol{\theta}} g(\boldsymbol{r}|\boldsymbol{\theta}',\alpha',\boldsymbol{y})\omega_m(\alpha')d\alpha' d\alpha d\boldsymbol{r} \\
&= \int_{\mathcal{Q}}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta},\alpha|\boldsymbol{r},\boldsymbol{y})g^*(\boldsymbol{r}|\boldsymbol{\theta}',\alpha',\boldsymbol{y})\times \\
&\qquad \times \left[\frac{\omega_m(\alpha)}{\omega(\alpha)} \frac{\int_\Theta\int_{\mathbb{R}_+} g(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{\theta},\alpha)g(\boldsymbol{r}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta})\omega(\alpha)d\alpha d\boldsymbol{\theta}}{\int_\Theta\int_{\mathbb{R}_+} g(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{\theta},\alpha)g(\boldsymbol{r}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta})\omega_m(\alpha)d\alpha d\boldsymbol{\theta}}\right] \omega_m(\alpha')d\alpha' d\alpha d\boldsymbol{r}.
\end{aligned}
$$

Thus, it is clear that the assumptions of the theorem do not concern the convergence of the integral over $\mathcal{Q}$, but only the convergence of its integrand. Assuming that $\omega_m(\alpha) \to \omega(\alpha)$ as $m \to \infty$ (which van Dyk and Meng never say in their theorem) in conjunction with the assumptions of Theorem 3.2 can only be used to conclude that the limit of the quantity in the square brackets is 1. Even under some regularity conditions, that would allow us to pass the limit inside the integral, the only thing that can be concluded is that

$$
k_m(\boldsymbol{\theta}|\boldsymbol{\theta}') \to \int_{\mathcal{Q}}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+} g^*(\boldsymbol{\theta},\alpha|\boldsymbol{r},\boldsymbol{y})g^*(\boldsymbol{r}|\boldsymbol{\theta}',\alpha',\boldsymbol{y})\omega(\alpha')d\alpha' d\alpha d\boldsymbol{r} = \infty \text{ as } m \to \infty.
$$

Meng and van Dyk (1999) provide other "checkable" and even more obscure conditions (Lemma 3 and Theorem 2) "ensuring" that Φ^*_{MA} is a positive recurrent Markov chain with $f(\boldsymbol{\theta}|\boldsymbol{y})$ as its unique invariant density. However, we do not pursue them any further. Instead, we will point out to another problem with their Lemma 1.

3.6 Other Problems with van Dyk and Meng's Lemma 1

We conclude this chapter with an interesting question that arises when working with sequences of random vectors. The question is whether the convergence of the joint densities implies convergence of the corresponding marginal densities. Here we show that in general this is not true and give an example to illustrate that. The reason we are interested in this question is because in van Dyk and

Meng's (2001) "proof" of their Lemma 1 on page 10, the authors claim that the convergence of $g_m(\boldsymbol{\theta}|\boldsymbol{r},\boldsymbol{y})$ to $g(\boldsymbol{\theta}|\boldsymbol{r},\boldsymbol{y})$ follows from their assumption that $g_m(\boldsymbol{\theta},\alpha|\boldsymbol{r},\boldsymbol{y}) \to g(\boldsymbol{\theta},\alpha|\boldsymbol{r},\boldsymbol{y})$ as $m \to \infty$. Here we show that this is not true.

Lemma 3.3. *Let $(X_n, Y_n), n = 1, 2, \ldots$ be a sequence of random vectors with values in $\mathscr{X} \times \mathscr{Y}$ and densities $f_n(x,y), n = 1, 2, \ldots$ (with respect to the Lebesgue measure). Suppose that*

$$f_n(x,y) \to f(x,y),\ \forall\ (x,y) \in \mathscr{X} \times \mathscr{Y} \text{ as } n \to \infty,$$

where $f(x,y)$ is a density on $\mathscr{X} \times \mathscr{Y}$. Then

$$\liminf_{n\to\infty} \int_{\mathscr{Y}} f_n(x,y)dy = \int_{\mathscr{Y}} f(x,y)dy \quad a.s.$$

Proof. By Fatou's lemma we have

$$\liminf_{n\to\infty} \int_{\mathscr{Y}} f_n(x,y)dy \geq \int_{\mathscr{Y}} \liminf_{n\to\infty} f_n(x,y)dy = \int_{\mathscr{Y}} f(x,y)dy, \tag{3.26}$$

for any $x \in \mathscr{X}$. Applying Fatou's lemma again we obtain

$$\int_{\mathscr{X}} \liminf_{n\to\infty} \int_{\mathscr{Y}} f_n(x,y)dydx \leq \liminf_{n\to\infty} \int_{\mathscr{X}} \int_{\mathscr{Y}} f_n(x,y)dydx = 1.$$

On the other hand, from (3.26) it follows

$$\int_{\mathscr{X}} \liminf_{n\to\infty} \int_{\mathscr{Y}} f_n(x,y)dydx \geq \int_{\mathscr{X}} \int_{\mathscr{Y}} f(x,y)dydx = 1,$$

and thus,

$$\begin{aligned} &\int_{\mathscr{X}} \liminf_{n\to\infty} \int_{\mathscr{Y}} f_n(x,y)dydx = 1 \\ \iff &\int_{\mathscr{X}} \left[\liminf_{n\to\infty} \int_{\mathscr{Y}} f_n(x,y)dy - \int_{\mathscr{Y}} f(x,y)dy\right] dx = 0. \end{aligned}$$

But by (3.26) the integrand is a non-negative function and therefore we must have

$$\liminf_{n\to\infty} \int_{\mathscr{Y}} f_n(x,y)dy = \int_{\mathscr{Y}} f(x,y)dy \text{ a.s.}$$

□

The following example shows that without further assumptions, there is no guarantee that the marginal densities even converge, much less that they converge to the desired limit.

Example 3.3: Let $n \in \mathbb{N} \setminus \{1\}$ and for $k \in \mathbb{N}_n$ define the following sequence of bivariate densities:

$$\tilde{f}_{nk}(x,y) = n^{\frac{3}{2}} I_{B_{nk}}(x,y) + \frac{n^2 - n^{\frac{3}{2}}}{n^2 - 1} I_{B_{11} \setminus B_{nk}}(x,y),$$

where

$$B_{nk} = \left\{ (x,y) \in \mathbb{R}^2 : \frac{k-1}{n} < x \leq \frac{k}{n}, 0 < y \leq \frac{1}{n} \right\}.$$

For $n = 1$ simply take $\tilde{f}_{11}(x,y) = I_{B_{11}}(x,y)$ which is the uniform density on the unit square.

Now consider the sequence $f_1(x,y), f_2(x,y), f_3(x,y), f_4(x,y), \ldots$ which is obtained by re-enumerating $\tilde{f}_{11}(x,y), \tilde{f}_{21}(x,y), \tilde{f}_{22}(x,y), \tilde{f}_{31}(x,y), \ldots$. Since for any particular $(x,y) \in B_{11}$ eventually as n gets large, $(x,y) \in B_{11} \setminus B_{nk}$, $\forall\, k = 1, \ldots, n$, it is clear that

$$f_n(x,y) \to \tilde{f}_{11}(x,y), \ \forall\ (x,y) \in B_{11} \text{ as } n \to \infty.$$

On the other hand the marginal densities

$$\tilde{g}_{nk}(x) = \int_0^1 \tilde{f}_{nk}(x,y)dy = \frac{n^{\frac{3}{2}} + n}{n+1} I_{(\frac{k-1}{n}, \frac{k}{n}]}(x) + \frac{n^2 - n^{\frac{3}{2}}}{n^2 - 1} I_{(0,1] \setminus (\frac{k-1}{n}, \frac{k}{n}]}(x)$$

do not converge to any version of the uniform density on the unit interval. Moreover, $\limsup_{n\to\infty} g_n(x) = \infty$ for all $0 < x \leq 1$ even though $\liminf_{n\to\infty} g_n(x) = I_{(0,1]}(x)$ as it is expected to be based on Lemma 1, where the sequence $\{g_n(x)\}$ is obtained from the sequence $\{g_{nk}(x)\}$ the same way as $\{f_n(x,y)\}$ from $\{f_{nk}(x,y)\}$. To see this, let $x \in (0,1]$ and notice that for every $n = 1, 2, \ldots$ there is a $k \in \{1, \ldots, n\}$ such that $\frac{k-1}{n} < x \leq \frac{k}{n}$ and this k is given by $k = \lceil nx \rceil$. Therefore

the subsequence $\tilde{g}_{n\lceil nx \rceil}(x) \to \infty$ as $n \to \infty$. The result about $\liminf_{n\to\infty} g_n(x)$ is obtained similarly.

The question we answer in the next chapter is "How fast do Φ^*_{MA} for the multivariate Student's t problem converge?" The Markov chain theory described in Section 4.2 will allow us to formulate and then answer a more precise version of this question.

CHAPTER 4
GEOMETRIC ERGODICITY OF Φ^*_{MA} FOR THE MULTIVARIATE STUDENT'S t PROBLEM

4.1 Introduction

Meng and van Dyk (1999) suggested that MA with an improper prior could be applied to the univariate version of the Student's t problem ($d = 1$), and van Dyk and Meng (2001) extended this to the multivariate case. These authors provided considerable empirical evidence suggesting that this new algorithm converges much faster than the standard data augmentation algorithm. However, they stopped short of formally proving anything about the convergence rate of the underlying marginal chain. In a discussion of van Dyk and Meng (2001), Hobert (2001a) showed that when $d = 1$ and $n = 2$, the marginal chain converges to stationarity at a geometric rate; that is, the chain is *geometrically ergodic* (Meyn and Tweedie, 1993, Chapter 15). In this chapter, we extend Hobert's results to the case $d \geq 1$ and $n \geq 2$. This is the first general, rigorous analysis of an MCMC algorithm based on a non-positive recurrent Markov chain.

There are two important *practical* applications of the results established in this chapter. First, if geometric ergodicity is established through *drift* and *minorization* conditions, as it is here, then formulas like those of Rosenthal (1995) and Roberts and Tweedie (1999) can be used to calculate an appropriate *burn-in* before sampling begins. In other words, given a starting value for the Markov chain, one can calculate an exact upper bound on the number of iterations required to get the total variation distance to stationarity below a pre-specified level. Second, geometric ergodicity implies the existence of central limit theorems that can be used to assess the Monte Carlo error of estimates of posterior quantities

of interest. This assessment can be carried out in a number of ways including regenerative simulation (Mykland, Tierney and Yu, 1995; Hobert, Jones, Presnell and Rosenthal, 2002) and time series-type methods (Geyer, 1992; Chan and Geyer, 1994).

The rest of this chapter is organized as follows. The necessary Markov chain background is provided in Section 4.2. The main results are given in Section 4.3 and they are applied to a simple illustrative example in Section 4.4. Finally, some concluding remarks are given in Section 4.5.

4.2 Markov Chain Background

In this section we give some basic definitions and then briefly describe what drift and minorization conditions are and how they can be used to establish geometric convergence of Markov chains. For a more general discussion of the ideas described in this section, see Jones and Hobert (2001). Let

$$X = \{X_i\}_{i=0}^{\infty}$$

be a time-homogeneous discrete-time Markov chain defined on the probability space $(\Omega, \mathscr{F}, \Pr)$ that takes values in the measurable state space $(\mathcal{X}, \mathscr{B})$. For $t \in \mathbb{N}$, let $P^t(x, A)$ be the t-step transition kernel [1] for X; that is, for all $x \in \mathcal{X}$ and all $A \in \mathscr{B}$,

$$P^t(x, A) = \Pr(X_{s+t} \in A | X_s = x) \ \ \forall\, s \in \{0, 1, 2, \ldots\}.$$

The following definitions can be found, for example, in Nummelin (1984) or Meyn and Tweedie (1993), but we state them here for completeness of the exposition.

[1] The superscript will be omitted when $t = 1$

Definition 4.1. *The Markov chain X is called γ-irreducible if there exists a measure γ on $\mathscr{B}$ such that, whenever $\gamma(A) > 0$, we have*

$$L(x, A) > 0 \ \ \forall\, x \in \mathcal{X},$$

where

$$L(x, A) := Pr(X \text{ ever enters } A | X_0 = x).$$

Definition 4.2. *If π is a σ-finite measure on $\mathscr{B}$ such that*

$$\pi(A) = \int_{\mathcal{X}} P(x, A)\pi(dx) \quad \forall\, A \in \mathscr{B},$$

then π is called an invariant measure for the Markov chain X.

Definition 4.3. *The π-irreducible Markov chain X is periodic if there exist an integer $d \geq 2$ and a collection $\{A_0, A_1, \ldots, A_{d-1}\}$ of nonempty and disjoint subsets in $\mathscr{B}$ such that, for all $i = 0, \ldots, d-1$ and all $x \in A_i$,*

$$P(x, A_j) = 1 \ \text{ for } j = i + 1 \mod d.$$

Otherwise, X is called aperiodic.

For the rest of the section, let $\mathscr{B}_+ := \{A \in \mathscr{B} : \pi(A) > 0\}$.

Definition 4.4. *The Markov chain X is called Harris recurrent if it is π-irreducible and for every set $A \in \mathscr{B}_+$,*

$$Pr(\eta_A = \infty | X_0 = x) = 1 \quad \forall\, x \in A,$$

where the occupation time random variable, $\eta_A := \sum_{n=1}^{\infty} I_A(X_n)$, counts the number of visits to the set A.

We will say that a Markov chain satisfies the *usual regularity conditions* if it:

(a) possesses an invariant probability measure π;

(b) is π-irreducible;

(c) is aperiodic; and

(d) is Harris recurrent.

Under the usual regularity conditions, for every $x \in \mathcal{X}$, we have the following limit for the total variation distance between $P^t(x, \cdot)$ and $\pi(\cdot)$:

$$\sup_{A \in \mathscr{B}} \left| P^t(x, A) - \pi(A) \right| =: \| P^t(x, \cdot) - \pi(\cdot) \| \to 0 \quad \text{as} \quad t \to \infty \, .$$

If this convergence takes place at a geometric rate; that is, if there exist a constant $0 \leq \phi < 1$ and a function $M : \mathcal{X} \to [0, \infty)$ such that for any $x \in \mathcal{X}$,

$$\| P^t(x, \cdot) - \pi(\cdot) \| \leq M(x) \, \phi^t$$

for all $t \in \mathbb{N}$, then X is said to be *geometrically ergodic*.

One method of proving that X is geometrically ergodic and, simultaneously, getting an upper bound on $M(x)\,\phi^t$ is by establishing *drift* and *minorization* conditions. There are several ways to do this (Meyn and Tweedie, 1994; Rosenthal, 1995; Roberts and Tweedie, 1999). The version we describe here is based on the work of Rosenthal (1995). A drift condition holds if for some function

$$V : \mathcal{X} \to [0, \infty),$$

and some constants $\lambda \in (0, 1)$ and $L \in \mathbb{R}$

$$PV(x) \leq \lambda V(x) + L \quad \forall \; x \in \mathcal{X} \, ,$$

where

$$PV(x) := \int_{\mathcal{X}} V(y) P(x, dy).$$

An associated minorization condition holds if for some probability measure H on $\mathscr{B}$ and some $\varepsilon > 0$

$$P(x, A) \geq \varepsilon H(A) \quad \forall \, x \in S \quad \forall \, A \in \mathscr{B},$$

where $S := \{x \in \mathcal{X} : V(x) \leq l\}$ with l being any number larger than $2L/(1 - \lambda)$. Together, drift and minorization imply that X is geometrically ergodic, and once

drift and minorization conditions are established, Theorem 12 from Rosenthal (1995) gives a formula for calculating an upper bound on $M(x)\,\phi^t$. This bound can be used to compute a sufficiently large burn-in (see Section 4.4 for an example).

4.3 Multivariate Student's t Problem

We now establish drift and minorization (and hence geometric ergodicity) for van Dyk and Meng's (2001) MA algorithm for the general multivariate Student's t problem described at the beginning of Section 2.1. The algorithm is based on the multivariate analogue of (3.19). Conditional on $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)$, let $(\boldsymbol{y}_j, q_j)$, $j \in \mathbb{N}_n$ be iid pairs satisfying the relations

$$\begin{aligned} \boldsymbol{y}_j | q_j, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha &\sim \mathrm{N}_d\left(\boldsymbol{\mu}, \frac{\alpha \boldsymbol{\Sigma}}{q_j}\right) \\ q_j | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha &\sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2\alpha}\right). \end{aligned} \tag{4.1}$$

We take the prior on $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)$ to be $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,\omega(\alpha)$, where $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-\frac{d+1}{2}}$ and $\omega(\alpha) = \alpha^{-1} I_{(0,\infty)}(\alpha)$. Let

$$g^*(\boldsymbol{q} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \boldsymbol{y}) := \frac{g^*(\boldsymbol{q}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha | \boldsymbol{y})}{\int_{\mathbb{R}_+^n} g^*(\boldsymbol{q}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha | \boldsymbol{y}) d\boldsymbol{q}}, \tag{4.2}$$

and

$$g^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha | \boldsymbol{q}, \boldsymbol{y}) := \frac{g^*(\boldsymbol{q}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha | \boldsymbol{y})}{\int_{\mathbb{R}_+} \int_S \int_{\mathbb{R}^d} g^*(\boldsymbol{q}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha | \boldsymbol{y}) d\boldsymbol{\mu} d\boldsymbol{\Sigma} d\alpha} \tag{4.3}$$

be the "conditional densities" based on (4.1). As in Subsection 3.3, these are legitimate densities despite the fact that the posterior $g^*(\boldsymbol{q}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha | \boldsymbol{y})$, based on (4.1), is improper. As a matter of fact we can even derive the explicit distribution of (4.2) and (4.3). First,

$$q_j | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \boldsymbol{y} \sim \Gamma\left(\frac{\nu + d}{2}, \frac{(\boldsymbol{y}_j - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}) + \nu}{2\alpha}\right),$$

independently for $j = 1, \ldots, n$. Furthermore, it is easy to identify all the conditionals in the decomposition

$$g^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha | \boldsymbol{q}, \boldsymbol{y}) = g^*(\boldsymbol{\mu} | \boldsymbol{\Sigma}, \alpha, \boldsymbol{q}, \boldsymbol{y})\, g^*(\boldsymbol{\Sigma} | \alpha, \boldsymbol{q}, \boldsymbol{y})\, g^*(\alpha | \boldsymbol{q}, \boldsymbol{y})\,.$$

Indeed,

$$\begin{aligned} \boldsymbol{\mu} | \boldsymbol{\Sigma}, \alpha, \boldsymbol{q}, \boldsymbol{y} &\sim \mathrm{N}_d \left(\hat{\boldsymbol{\mu}}, \frac{\alpha \boldsymbol{\Sigma}}{q_\cdot} \right) \\ \boldsymbol{\Sigma} | \alpha, \boldsymbol{q}, \boldsymbol{y} &\sim \mathrm{IW}_d \left[n - 1, \alpha \left(\sum_{j=1}^{n} q_j (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})^\top \right)^{-1} \right] \end{aligned}$$

and

$$\alpha | \boldsymbol{q}, \boldsymbol{y} \sim \mathrm{IG} \left(\frac{n\nu}{2}, \frac{\nu q_\cdot}{2} \right),$$

where $\hat{\boldsymbol{\mu}} = q_\cdot^{-1} \sum_{j=1}^{n} q_j \boldsymbol{y}_j$ and by $\boldsymbol{X} \sim \mathrm{IW}_p(k, \boldsymbol{\Lambda})$ we mean that the $p \times p$ random matrix $\boldsymbol{X}$ has Inverse Wishart distribution (see; e.g. Mardia, Kent and Bibby, 1979) with density

$$f_{\mathrm{IW}_p}(\boldsymbol{X}; k, \boldsymbol{\Lambda}) = \frac{|\boldsymbol{X}|^{-\frac{k+p+1}{2}} e^{-\frac{\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{X}^{-1})}{2}}}{2^{\frac{kp}{2}} \pi^{\frac{p(p-1)}{4}} |\Lambda|^{\frac{k}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(k+1-i)\right)}.$$

It is straightforward to show that everything from the univariate case, described in Example 3.1, follows through; that is, we can use the conditionals in (4.2) and (4.3) to construct a non-positive recurrent two-variable Gibbs sampler that has a marginal chain

$$\Phi^*_{\mathrm{MA}} = \{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}_{i=0,1,2,\ldots}\,,$$

which satisfies the usual regularity conditions and has (2.1) as its invariant density. The Markov transition density of Φ^*_{MA} is

$$K_{\boldsymbol{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{\mu}', \boldsymbol{\Sigma}') = \int_{\mathbb{R}_+} \int_{\mathbb{R}^q_+} g^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha | \boldsymbol{q}, \boldsymbol{y})\, g^*(\boldsymbol{q} | \boldsymbol{\mu}', \boldsymbol{\Sigma}', \alpha', \boldsymbol{y})\, d\boldsymbol{q}\, d\alpha\,,$$

where α' can be taken to be any positive number.

We now prove a lemma that will be needed in establishing the drift condition.

Lemma 4.1. *Let $\boldsymbol{A}$ be an $m \times m$ positive definite matrix. Then*

$$\boldsymbol{x}^\top(\boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{A})^{-1}\boldsymbol{x} < 1 \quad \forall\, \boldsymbol{x} \in \mathbb{R}^m.$$

Proof. Applying the Sherman-Morrison-Woodbury formula (see, for example, Golub and Van Loan (1989), page 51) we obtain

$$(\boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{A})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{x}(1 + \boldsymbol{x}^\top\boldsymbol{A}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^\top\boldsymbol{A}^{-1}.$$

Therefore,

$$\boldsymbol{x}^\top(\boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{A})^{-1}\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{A}^{-1}\boldsymbol{x} - \frac{\boldsymbol{x}^\top\boldsymbol{A}^{-1}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{A}^{-1}\boldsymbol{x}}{1 + \boldsymbol{x}^\top\boldsymbol{A}^{-1}\boldsymbol{x}} = t - \frac{t^2}{1+t} = \frac{t}{1+t} < 1,$$

where $t = \boldsymbol{x}^\top\boldsymbol{A}^{-1}\boldsymbol{x} \geq 0$, since $\boldsymbol{A}$ is positive definite. □

The following result yields the drift condition:

Theorem 4.1. *Define*

$$V(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}).$$

If $n\nu > 2$ and $\nu + d > 2$ then

$$PV(\boldsymbol{\mu}', \boldsymbol{\Sigma}') \leq \frac{(n-1)^2(\nu+d)}{(n\nu-2)(\nu+d-2)} V(\boldsymbol{\mu}', \boldsymbol{\Sigma}') + \frac{nd\nu}{n\nu-2} + \frac{n\nu(n-1)(n\nu+nd-2)}{(n\nu-2)(\nu+d-2)}.$$

Proof. We will evaluate the expectation of the drift function using several layers of iterated expectations:

$$\begin{aligned} PV(\boldsymbol{\mu}', \boldsymbol{\Sigma}') &= \mathrm{E}\big[\mathrm{E}(V(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\boldsymbol{q}, \boldsymbol{y})|\boldsymbol{\mu}', \boldsymbol{\Sigma}', \boldsymbol{y}\big] \\ &= \mathrm{E}\Big\{\mathrm{E}\big[\mathrm{E}(V(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\alpha, \boldsymbol{q}, \boldsymbol{y})|\boldsymbol{q}, \boldsymbol{y}\big]\big|\boldsymbol{\mu}', \boldsymbol{\Sigma}', \boldsymbol{y}\Big\} \\ &= \mathrm{E}\left(\mathrm{E}\Big\{\mathrm{E}\big[\mathrm{E}(V(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\boldsymbol{\Sigma}, \alpha, \boldsymbol{q}, \boldsymbol{y})|\alpha, \boldsymbol{q}, \boldsymbol{y}\big]\big|\boldsymbol{q}, \boldsymbol{y}\Big\}\Big|\boldsymbol{\mu}', \boldsymbol{\Sigma}', \boldsymbol{y}\right). \end{aligned}$$

The expectation on the left-hand side is with respect to the Markov transition density $K_{\boldsymbol{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\mu}', \boldsymbol{\Sigma}')$ and the first equality follows from the properties of the

Gibbs sampler. The other two equalities are nothing more than the usual iterated expectation.

Starting with the innermost expectation, we obtain

$$
\begin{aligned}
&\mathrm{E}\left[V(\boldsymbol{\mu},\boldsymbol{\Sigma})|\boldsymbol{\Sigma},\alpha,\boldsymbol{q},\boldsymbol{y}\right] \\
&= \sum_{i=1}^{n} \mathrm{E}[(\boldsymbol{y}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu})|\boldsymbol{\Sigma},\alpha,\boldsymbol{q},\boldsymbol{y}] \\
&= \sum_{i=1}^{n} \mathrm{E}[\boldsymbol{y}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_i - \boldsymbol{y}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_i + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}|\boldsymbol{\Sigma},\alpha,\boldsymbol{q},\boldsymbol{y}] \\
&= \sum_{i=1}^{n} \boldsymbol{y}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_i - \sum_{i=1}^{n} \boldsymbol{y}_i^\top\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}^\top\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\boldsymbol{y}_i + n\left[\hat{\boldsymbol{\mu}}^\top\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}} + \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\alpha\boldsymbol{\Sigma}}{q_\cdot}\right)\right] \\
&= \sum_{i=1}^{n} (\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}) + \frac{nd\alpha}{q_\cdot}\,.
\end{aligned}
$$

Now if $\boldsymbol{X} \sim \mathrm{IW}_p(k,\boldsymbol{\Lambda})$, then $\boldsymbol{X}^{-1} \sim \mathrm{Wishart}_p(k,\boldsymbol{\Lambda})$ and $\mathrm{E}\boldsymbol{X}^{-1} = k\boldsymbol{\Lambda}$ (see, for example, Mardia et al. (1979), page 85). Thus,

$$
\begin{aligned}
\mathrm{E}\left[V(\boldsymbol{\mu},\boldsymbol{\Sigma})|\alpha,\boldsymbol{q},\boldsymbol{y}\right] &= \frac{nd\alpha}{q_\cdot} + \sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})^\top\mathrm{E}\left(\boldsymbol{\Sigma}^{-1}|\alpha,\boldsymbol{q},\boldsymbol{y}\right)(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}) \\
&= \frac{nd\alpha}{q_\cdot} + \alpha(n-1)\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})^\top\left[\sum_{j=1}^{n} q_j(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})^\top\right]^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}).
\end{aligned}
$$

Since $\nu n > 2$, conditional on $(\boldsymbol{q},\boldsymbol{y})$, α has a finite mean given by $\mathrm{E}(\alpha|\boldsymbol{q},\boldsymbol{y}) = \frac{\nu q_\cdot}{\nu n-2}$. Therefore

$$
\begin{aligned}
&\mathrm{E}\left[V(\boldsymbol{\mu},\boldsymbol{\Sigma})|\boldsymbol{q},\boldsymbol{y}\right] = \frac{nd}{q_\cdot}\mathrm{E}(\alpha|\boldsymbol{q},\boldsymbol{y}) \\
&+ (n-1)\mathrm{E}(\alpha|\boldsymbol{q},\boldsymbol{y})\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})^\top\left[\sum_{j=1}^{n} q_j(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})^\top\right]^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}) \qquad (4.4) \\
&= \frac{nd\nu}{n\nu-2} + \frac{\nu(n-1)}{n\nu-2}q_\cdot\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})^\top\left[\sum_{j=1}^{n} q_j(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})^\top\right]^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}).
\end{aligned}
$$

Now, we need a bound for the quantity

$$
q_\cdot\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})^\top\left[\sum_{j=1}^{n} q_j(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})^\top\right]^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})\,, \qquad (4.5)
$$

which is obtained by applying Lemma 4.1 to the ith component in (4.5) as follows

$$(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})^\top \left[\sum_{j=1}^{n} q_j (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})^\top \right]^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})$$

$$= (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})^\top \frac{1}{q_i} \left[\sum_{j \neq i}^{n} \frac{q_j}{q_i} (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})^\top + (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})^\top \right]^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}) \leq \frac{1}{q_i}.$$

Therefore from (4.4) we obtain that

$$\mathrm{E}[V(\boldsymbol{\mu}, \boldsymbol{\Sigma}) | \boldsymbol{q}, \boldsymbol{y}] \leq \frac{nd\nu}{n\nu - 2} + \frac{\nu(n-1)}{n\nu - 2} q_\cdot \sum_{i=1}^{n} \frac{1}{q_i} = \frac{nd\nu}{n\nu - 2} + \frac{\nu(n-1)}{n\nu - 2} \left(n + \sum_{i=1}^{n} \sum_{j \neq i} \frac{q_j}{q_i} \right).$$

Finally,

$$\mathrm{E}_{\boldsymbol{y}}[V(\boldsymbol{\mu}, \boldsymbol{\Sigma}) | \boldsymbol{\mu}', \boldsymbol{\Sigma}']$$

$$\leq \frac{nd\nu}{n\nu - 2} + \frac{\nu(n-1)}{n\nu - 2} \left[\frac{\nu + d}{\nu + d - 2} \sum_{i=1}^{n} \sum_{j \neq i} \frac{(\boldsymbol{y}_i - \boldsymbol{\mu}')^\top (\boldsymbol{\Sigma}')^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}') + \nu}{(\boldsymbol{y}_j - \boldsymbol{\mu}')^\top (\boldsymbol{\Sigma}')^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}') + \nu} + n \right]$$

$$\leq \frac{nd\nu}{n\nu - 2} + \frac{\nu(n-1)}{n\nu - 2} \left[\frac{(\nu + d)(n-1)}{\nu(\nu + d - 2)} \sum_{i=1}^{n} \left((\boldsymbol{y}_i - \boldsymbol{\mu}')^\top (\boldsymbol{\Sigma}')^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}') + \nu \right) + n \right],$$

and some rearrangement yields the result. □

Remark 4.1. *This result agrees perfectly with Hobert's (2001a) result in the case where $d = 1$ and $n = 2$.*

Remark 4.2. *A similar result was obtained by Jones (2001) for the case $d = 1$, but our result is uniformly better in the sense that our coefficient in front of $V(\mu', \sigma')$ is always smaller and doesn't depend on the data.*

The last thing we need is the minorization condition. The following is a straightforward generalization of the minorization developed in Hobert (2001a).

Theorem 4.2. *Fix $l > 0$ and let $S = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : V(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq l\}$. Then the Markov transition density $K_{\boldsymbol{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{\mu}', \boldsymbol{\Sigma}')$ satisfies the following minorization condition*

$$K_{\boldsymbol{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{\mu}', \boldsymbol{\Sigma}') \geq \varepsilon h(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{y}) \quad \forall \ (\boldsymbol{\mu}', \boldsymbol{\Sigma}') \in S,$$

where the density $h(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y})$ *is given by*

$$h(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y}) = \int_0^\infty \int_{\mathbb{R}_+^n} g^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha|\boldsymbol{q}, \boldsymbol{y}) \left[\prod_{i=1}^n \frac{g(q_i)}{\int_0^\infty g(t)\, dt} \right] d\boldsymbol{q}\, d\alpha \,,$$

$\varepsilon = \left(\int_0^\infty g(t)\, dt\right)^n$ *and the function* $g(\cdot)$ *is given by*

$$g(q) = \Gamma\left(\frac{\nu+d}{2}, \frac{\nu}{2}; q\right) I_{(0,q^*)}(q) + \Gamma\left(\frac{\nu+d}{2}, \frac{\nu+l}{2}; q\right) I_{(q^*,\infty)}(q),$$

where $\Gamma(a, b; x)$ *denotes the Gamma density with parameters* a *and* b *evaluated at the point* x *and* $q^* = \frac{\nu+d}{l} \log\left(1 + \frac{l}{\nu}\right)$.

Together, Theorems 4.1 and 4.2 show that for many (d, ν, n) triples, Φ^*_{MA} is geometrically ergodic. We state this as a corollary.

Corollary 4.1. *The Markov chain* Φ^*_{MA} *is geometrically ergodic if* $n\nu > 2$, $\nu + d > 2$ *and*

$$\frac{(n-1)^2(\nu+d)}{(n\nu-2)(\nu+d-2)} < 1\,. \tag{4.6}$$

When $\nu \geq 2$, condition (4.6) is satisfied for

$$n < \frac{\nu^2 + \nu d + 2d + \sqrt{(\nu+d-2)[d(\nu^2+4\nu-8) + \nu(\nu+4)(\nu-2)]}}{2(\nu+d)}.$$

Roughly speaking, all the conditions of Corollary 4.1 hold when $n \leq \nu$ if $d \geq 2$ and when $n \leq \nu - 1$ if $d = 1$.

Of course, ideally, we would like to be able to say that Φ^*_{MA} is geometrically ergodic for all (d, ν, n) triples such that $d \geq 1$, $\nu > 0$ and $n \geq d + 1$. The most restrictive of the three conditions in Corollary 4.1 is clearly (4.6) which is a direct result of the upper bound we use for the quantity (4.5) in the proof of Theorem 4.1. (Note that up until this point in the proof, all of the expectations are evaluated *exactly*.) Our experience suggests that finding a closed form expression for the expectation of (4.5) is impossible. Furthermore, alternative upper bounds that one might consider using are either not as sharp or have other undesirable consequences. For example, note that our bound on (4.5) holds uniformly in $\boldsymbol{y}$. If

a bound depending on $\boldsymbol{y}$ is utilized, then the coefficient in front of $V(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$ in the drift condition will also depend on $\boldsymbol{y}$. This could lead to the unfortunate situation where the sufficient conditions for geometric convergence of the Markov chain actually depend on the data. It is our belief that a substantial improvement on Theorem 4.1 will likely require the use of a different drift function, $V(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which basically means starting over from square one.

Amazingly, the standard DA algorithm is also geometrically ergodic. We prove this fact in the next theorem which is an analog of Theorem 4.1.

Theorem 4.3. *Let $V(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be defined as in Theorem 4.1. If $\nu + d > 2$ then*

$$E_{\boldsymbol{y}}\left[V(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\boldsymbol{\mu}', \boldsymbol{\Sigma}'\right] \leq \frac{n+d-1}{\nu+d-2} V(\boldsymbol{\mu}', \boldsymbol{\Sigma}') + \frac{n(n+d-1)\nu}{\nu+d-2},$$

where $E_{\boldsymbol{y}}[\cdot|\cdot]$ is with respect to the Markov transition density of the standard DA chain.

Proof. It is easy to check that for the standard DA we have the following facts. First,

$$q_j|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{y} \sim \Gamma\left(\frac{\nu+d}{2}, \frac{(\boldsymbol{y}_j - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}) + \nu}{2}\right),$$

independently for $j = 1, ..., n$. Second,

$$\boldsymbol{\mu}|\boldsymbol{\Sigma}, \boldsymbol{q}, \boldsymbol{y} \sim \mathrm{N}_d\left(\hat{\boldsymbol{\mu}}, \frac{\boldsymbol{\Sigma}}{q_\cdot}\right),$$

and

$$\boldsymbol{\Sigma}|\boldsymbol{q}, \boldsymbol{y} \sim \mathrm{IW}_d\left[n-1, \left(\sum_{j=1}^{n} q_j(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})^\top\right)^{-1}\right].$$

Now we have that

$$\begin{aligned} \mathrm{E}_{\boldsymbol{y}}[V(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\boldsymbol{\mu}', \boldsymbol{\Sigma}'] &= \mathrm{E}\left[\mathrm{E}(V(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\boldsymbol{q}, \boldsymbol{y})|\boldsymbol{\mu}', \boldsymbol{\Sigma}', \boldsymbol{y}\right] \\ &= \mathrm{E}\Big\{\mathrm{E}\left[\mathrm{E}(V(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\boldsymbol{\Sigma}, \boldsymbol{q}, \boldsymbol{y})|\boldsymbol{q}, \boldsymbol{y}\right]|\boldsymbol{\mu}', \boldsymbol{\Sigma}', \boldsymbol{y}\Big\}. \end{aligned}$$

Skipping some of the details we obtain

$$\mathrm{E}\left[V(\boldsymbol{\mu},\boldsymbol{\Sigma})|\boldsymbol{q},\boldsymbol{y}\right] = \frac{nd}{q.} + (n-1)\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})^\top\left[\sum_{j=1}^{n}q_j(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}})^\top\right]^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})\,.$$

Now using the fact that $q_\cdot^{-1} \leq \frac{1}{n}\sum_{i=1}^{n} q_i^{-1}$ as well as the results established earlier in the proof of Theorem 4.1, we have

$$\mathrm{E}\left[V(\boldsymbol{\mu},\boldsymbol{\Sigma})|\boldsymbol{q},\boldsymbol{y}\right] \leq (n+d-1)\sum_{i=1}^{n}\frac{1}{q_i}\,.$$

Finally,

$$\mathrm{E}_{\boldsymbol{y}}[V(\boldsymbol{\mu},\boldsymbol{\Sigma})|\boldsymbol{\mu}',\boldsymbol{\Sigma}'] \leq \frac{n+d-1}{\nu+d-2}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i-\boldsymbol{\mu}')^\top(\boldsymbol{\Sigma}')^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}')+\nu\right],$$

and some rearrangement yields the result. □

The above theorem combined with the fact that Theorem 4.2 also applies to standard DA shows that the standard DA algorithm is also geometrically ergodic.

4.4 Numerical Example

Suppose that $d = 2$, $\nu = 10$, $n = 3$ and that the data are: $\boldsymbol{y}_1 = (-5,-2)^\top$, $\boldsymbol{y}_2 = (1,5)^\top$, $\boldsymbol{y}_3 = (-10,4)^\top$. In this section, we will use our results in conjunction with Theorem 12 from Rosenthal (1995) to calculate sufficient burn-in for Φ^*_{MA} and for the standard DA algorithm. We will also take advantage of the fact that it is possible to get iid samples from the target in this particular case (see Section 2.2) to examine empirically how long it takes for the Markov chains to converge.

For Φ^*_{MA}, Theorem 4.1 yields

$$\mathrm{E}_{\boldsymbol{y}}[V(\boldsymbol{\mu},\boldsymbol{\Sigma})|\boldsymbol{\mu}',\boldsymbol{\Sigma}'] \leq \lambda V(\boldsymbol{\mu}',\boldsymbol{\Sigma}') + L = \frac{6}{35}V(\boldsymbol{\mu}',\boldsymbol{\Sigma}') + \frac{66}{7}.$$

In order to apply Rosenthal's Theorem 12, we need to choose an $l > \frac{2L}{1-\lambda} \approx 22.76$ and an $r \in (0,1)$. Taking $l = 27.241$ and $r = 0.04$ leads to $q^* \approx 0.58$ and this yields $\varepsilon \approx 0.00158$. Finally, suppose we start the chain with $\boldsymbol{\mu}_0 = (0,0)^\top$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$.

Then our Theorems 4.1 and 4.2 in conjunction with Rosenthal's Theorem 12 yield

$$\|P_{\boldsymbol{y}}^t[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot] - \pi(\cdot|\boldsymbol{y})\| \leq (0.9999368199)^t + (183.3793103)(0.9996034577)^t ,$$

where $P_{\boldsymbol{y}}^t[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot]$ denotes the t-step Markov transition kernel corresponding to Φ_{MA}^* and $\pi(\cdot|\boldsymbol{y})$ denotes the probability measure corresponding to (2.1). It is easy to verify that

$$\|P_{\boldsymbol{y}}^{47415}[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot] - \pi(\cdot|\boldsymbol{y})\| < 0.05 .$$

Hence, a burn-in of 47,415 iterations for Φ_{MA}^* is more than enough. To put this in perspective, it takes about a minute to run 1,000,000 iterations of this Markov chain (on a slow laptop).

The reader may wonder how we chose these particular values for l and r. Because Rosenthal's bound holds for all $l > 2L/(1-\lambda)$ and all $r \in (0,1)$, the user is free to choose the values that lead to the smallest upper bound. In our applications, we simply used a grid-search to find the (approximate) minimizers.

For the standard DA algorithm, Theorem 4.3 yields

$$\mathrm{E}_{\boldsymbol{y}}[V(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\boldsymbol{\mu}', \boldsymbol{\Sigma}'] \leq \frac{2}{5} V(\boldsymbol{\mu}', \boldsymbol{\Sigma}') + 12 .$$

Taking $l = 46.017$ and $r = 0.019$ (and starting the chain at the same initial values as the MA chain) yields

$$\|R_{\boldsymbol{y}}^t[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot] - \pi(\cdot|\boldsymbol{y})\| \leq (0.999998668)^t + 23(0.9999831867)^t ,$$

where $R_{\boldsymbol{y}}^t[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot]$ is the t-step Markov transition kernel corresponding to the standard DA algorithm. In this case, we need 2,249,050 iterations to get the total variation distance below 0.05. That's nearly 50 times as many as are needed for Φ_{MA}^*!

It is tempting to conclude from this comparison that, in this particular case, Φ_{MA}^* converges faster than the Markov chain underlying standard DA . Keep in

mind, however, that we are comparing two *upper bounds*. It is still possible, albeit unlikely, that standard DA gets within 0.05 of the stationary distribution sooner than $\Phi^{\star}_{\mathrm{MA}}$. (See Jones and Hobert (2004) for some general discussion along these lines.)

Recall that n is exactly equal to $d+1$ in the example we are considering and hence it is possible to make iid draws from $\pi(\cdot|\boldsymbol{y})$ in this particular case. Thus, for fixed t, we can compare $\pi(\cdot|\boldsymbol{y})$, $P^t_{\boldsymbol{y}}[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot]$ and $R^t_{\boldsymbol{y}}[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot]$ empirically through random samples generated from the three distributions. A method of drawing a random sample from $\pi(\cdot|\boldsymbol{y})$ is given in Section 2.2. Making iid draws from $P^t_{\boldsymbol{y}}[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot]$ and $R^t_{\boldsymbol{y}}[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot]$ is simple - just run independent copies of the Markov chains (starting each at $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$) and take the tth iteration of each chain.

We simulated random samples of size 10^6 from $\pi(\cdot|\boldsymbol{y})$, $P^1_{\boldsymbol{y}}[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot]$, $P^3_{\boldsymbol{y}}[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot]$, and $P^{10}_{\boldsymbol{y}}[(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \cdot]$ as well as the corresponding distributions under standard DA. The 5-dimensional samples are compared on the basis of three univariate quantities: μ_1, σ_{11} (the first diagonal element of $\boldsymbol{\Sigma}$) and σ_{12} (the off-diagonal element of $\boldsymbol{\Sigma}$). The univariate comparisons are made via density estimates calculated using the `density()` command in the R programming language (under the default settings). The results are shown in Figure 4–1. These plots suggest that the standard DA algorithm takes a bit longer to converge than $\Phi^{\star}_{\mathrm{MA}}$ and that both algorithms have converged after about 10 iterations. Thus, the empirical results suggest that the burn-ins calculated above are extremely conservative. It should be kept in mind, however, that nothing conclusive about the total variation distance (between the 5-dimensional distributions) can be said based on such empirical comparisons.

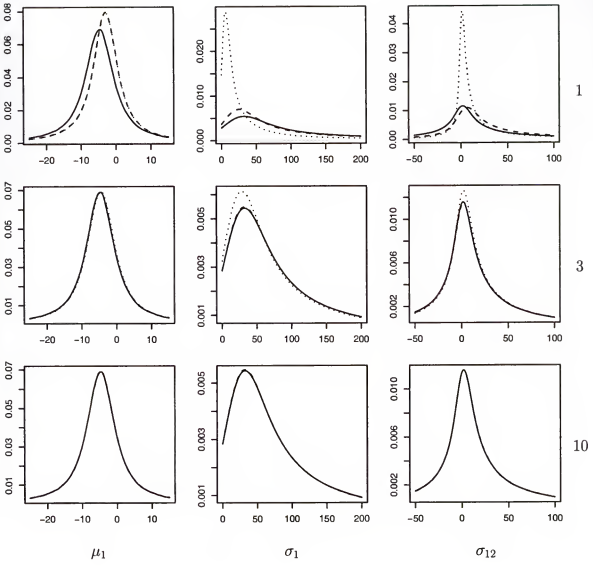

Figure 4–1: Empirical comparison of Markov chains and the target

From left to right, the three columns correspond to the parameters μ_1, σ_1 and σ_{12}. From top to bottom, the three rows correspond to 1 iteration, 3 iterations and 10 iterations. Each plot contains three density estimates and each density estimate is based on a random sample of size 10^6. The solid line, dashed line and dots correspond to the iid sample, Φ^*_{MA} and the standard DA algorithm, respectively. Since "iterations" are irrelevant for the iid sample, the same density estimate is shown in each column for the iid case.

4.5 Concluding Remarks

The results in Section 4.4 contribute to mounting evidence that upper bounds on the total variation distance calculated using Rosenthal's (1995) Theorem 12 (or the related results of Roberts and Tweedie (1999)) are typically very conservative. Indeed, Rosenthal (1995, Example 1) demonstrated such conservatism in the context of a simple Markov chain whose exact convergence rate is known. In the Rejoinder of van Dyk and Meng (2001), an empirical estimator of total variation distance was used to show that the amount of burn-in suggested by Hobert (2001a) (calculated using Rosenthal's (1995) Theorem 12) is probably much more than is actually needed. By applying both the techniques of Roberts and Tweedie (1999) and those of Rosenthal (1995) in the same example, Jones and Hobert (2004) illustrated the potential conservativeness of results based on Roberts and Tweedie's (1999) techniques.

On the other hand, suppose that one of these techniques is applied in a particular situation (where it is impossible to sample directly from the target) and the result is that n iterations is a sufficient burn-in. Even if we believe (say, because of the mounting evidence alluded to above) that n is probably much larger than is actually necessary, there is no harm in burning in n iterations, unless this is simply too time consuming. Thus, from a "better safe than sorry" standpoint, even an extremely conservative upper bound on the necessary amount of burn-in can be useful.

We conclude with a few comments regarding a possible generalization of our results. The basic DA algorithm underlying all of the results in this dissertation is based on the fact that the location-scale Student's t density can be written as a mixture of normals. For example, the d-variate Student's t density with ν degrees

of freedom, f_{t_ν}, can be written as

$$f_{t_\nu}(\boldsymbol{y}) = \int_{\mathbb{R}_+} (2\pi)^{-\frac{d}{2}} q^{\frac{d}{2}} e^{-\frac{q\boldsymbol{y}^T\boldsymbol{y}}{2}} \, dG(q) \,, \tag{4.7}$$

where G is the Gamma$(\nu/2, \nu/2)$ distribution function. Now consider a more general problem where the data come from a density having the same form as (4.7), but where G is the distribution function of any positive random variable. Bayesian regression models with errors of this form have been studied by many authors including West (1984), Liu (1996) and Fernández and Steel (1999, 2000). Perhaps general versions of Theorems 4.1 and 4.2 could be developed to handle this problem.

REFERENCES

Affleck-Grave, J. and McDonald, B. (1989). Nonnormalities and tests of assest pricing theories, *Journal of Finance* **44**: 889–908.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*, Wiley, New York.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, John Wiley, New York.

Chan, K. S. and Geyer, C. J. (1994). Comment on "Markov chains for exploring posterior distributions" by L. Tierney, *The Annals of Statistics* **22**: 1747–1758.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, John Wiley, New York.

Eaton, M. L. (1989). *Group Invariance Applications in Statistics*, Institute of Mathematical Statistics and the American Statistical Association, Hayward, California and Alexandria, Virginia.

Fang, K.-T. and Zhang, Y.-T. (1990). *Generalized Multivariate Analysis*, Springer-Verlag, Berlin.

Fernández, C. and Steel, M. F. J. (2000). Bayesian regression analysis with scale mixtures of normals, *Econometric Theory* **16**: 80–101.

Geweke, J. (1993). Bayesian treatment of the independent student-t linear model, *Journal of Applied Econometrics* **8**: S19–S40.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion), *Statistical Science* **7**: 473–511.

Golub, G. H. and Van Loan, C. F. (1989). *Matrix Computations*, The Johns Hopkins University Press, Baltimore and London.

Green, J. A. (1987). *Sets and Groups*, Routledge and Kegan Paul, London and New York.

Hobert, J. P. (2001a). Discussion of "The art of data augmentation" by D.A. van Dyk and X.-L. Meng, *Journal of Computational and Graphical Statistics* **10**: 59–68.

Hobert, J. P. (2001b). Stability relationships among the Gibbs sampler and its subchains, *Journal of Computational and Graphical Statistics* **10**: 185–205.

Hobert, J. P., Jones, G. L., Presnell, B. and Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo, *Biometrika* **89**: 731–743.

Jeffreys, H. (1939). *Theory of Probability*, Clarendon Press, Oxford, U.K.

Jones, G. L. (2001). *Convergence Rates and Monte Carlo Standard Errors for Markov Chain Monte Carlo Algorithms*, unpublished PhD dissertation, University of Florida, Gainesville.

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo, *Statistical Science* **16**: 312–334.

Jones, G. L. and Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model, *The Annals of Statistics* **32**: 784–817.

Karlin, S. (1968). *Total Positivity*, Vol. 1, Stanford University Press, Stanford, California.

Kelejian, H. H. and Prucha, I. R. (1985). Independent or uncorrelated disturbances in linear regression, *Economics Letter* **19**: 35–38.

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution, *Journal of the American Statistical Association* **84**: 881–896.

Lavine, M. (2003). A marginal ergodic theorem, *in* J. M. Bernardo, A. P. Dawid, J. O. Berger and M. West (eds), *Bayesian Statistics 7*, Oxford University Press, Oxford.

Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data, *Journal of the American Statistical Association* **91**: 1219–1227.

Liu, J. S., Wong, W. H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to comparisons of estimators and augmentation schemes, *Biometrika* **81**: 27–40.

Liu, J. S., Wong, W. H. and Kong, A. (1995). Covariance structure and convergence rate of the gibbs sampler with various scans, *Journal of the Royal Statistical Society, Series B, Methodological* **57**: 157–169.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation, *Journal of the American Statistical Association* **94**: 1264–1274.

Mandelbroit, B. (1963). The variation of certain speculative prices, *Journal of Business* **36**: 394–419.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, London.

Meng, X.-L. and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation, *Biometrika* **86**: 301–320.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*, Springer-Verlag, London.

Meyn, S. P. and Tweedie, R. L. (1994). Computable bounds for geometric convergence rates of Markov chains, *The Annals of Applied Probability* **4**: 981–1011.

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers, *Journal of the American Statistical Association* **90**: 233–241.

Nelson, C. R. and Plosser, C. I. (1982). Trends and random walks in macroeconomic time series: Some evidence and implications, *Journal of Monetary Economics* **10**: 139–162.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*, Cambridge, London.

Olver, P. J. (1999). *Classical Invariant Theory*, Cambridge University Press, Cambridge, United Kingdom.

Praetz, P. D. (1972). The distribution of share price changes, *The Journal of Business* **45**: 49–55.

Ripley, B. D. (1987). *Stochastic Simulation*, John Wiley and Sons, New York.

Roberts, G. and Tweedie, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains, *Stochastic Processes and their Applications* **80**: 211–229. Corrigendum (2001) **91**: 337-338.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo, *Journal of the American Statistical Association* **90**: 558–566.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion), *Journal of the American Statistical Association* **82**: 528–550.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *The Annals of Statistics* **22**: 1701–1762.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion), *Journal of Computational and Graphical Statistics* **10**: 1–50.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression, *Journal of the Royal Statistical Society,* Series B **46**: 431–439.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate student-t error terms, *Journal of the American Statistical Association* **71**: 400–405.

BIOGRAPHICAL SKETCH

I was born in Bulgaria, in 1974. My educational background consists of a high school diploma in Mathematics and a master's degree in Mathematics (with emphasis on Probability and Statistics) from Sofia University. Shortly after receiving my master's degree in 1997, I went to the Bulgarian army to fulfill a 9-month duty required by law. Immediately after that, in 1998, I was appointed as an Assistant Professor of Statistics in the Department of Economics and Business Administration at Sofia University. By that time I had already decided to pursue a Ph.D. degree in the USA. Luckily, I was accepted at the University of Florida's Department of Statistics, where I began studying in 1999. After graduating from the University of Florida, I will move to New York City to work as an Assistant Professor in the Department of Statistics in the Zicklin School of Business at Baruch College, City University of New York.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James P. Hobert, Chair
Associate Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Brett D. Presnell
Associate Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James G. Booth
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Adao A. Trindade
Assistant Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Murali Rao
Professor of Mathematics

This dissertation was submitted to the Graduate Faculty of the Department of Statistics in the College of Liberal Arts and Sciences and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 2004

Dean, Graduate School